Cda - H02959 - 18 - p 006778

Vol. 50 No. 1 (New Series No. 225) January 19

ANALYSIS

⑱

Edited by
PETER SMITH

CONTENTS

Liar syllogisms THEODORE M. DRAN
A variation on a paradox ALLEN HAZE
A defence of modus tollens
WALTER SINNOTT-ARMSTRONG, JAMES MOOR *and* ROBERT FOGEL
'Actually' ROGER TEICHMAN
Are Salmon's 'guises' disguised Fregean senses? JOÃO BRANQUINH
Honderich on the indispensability of the mental GREGORY MCCULLOC
Van Inwagen on the 'obviousness' of libertarian moral responsibility
SAUL SMILANSK
How to *Mind* one's *Ethics*: A reply to van Inwagen V. ALAN WHIT
Paternalism defined DAVID ARCHAR
Is the best really necessary? E. R. KRAEME
Intuitionism versus degrees of truth STEPHEN P. SCHWART
A doxastic paradox JON PEREZ LARAUDOGOITI

BASIL BLACKWELL · OXFORD AND NEW YORK

ISSN 0003-2638

ANALYSIS 50

Founded in 1933 and, apart from the war years, published continuously since, ANALYSIS has now reached its 50th volume. A Conference to mark and celebrate the occasion will be held at Newnham College, Cambridge, from Friday March 30 to Sunday April 1, 1990. Papers will be given by

- Jonathan Bennett, 'Why Beliefs Cannot be Voluntary'
- Aaron Ben-Zeev, 'Why did Psammenitus not Pity his Son?'
- Simon Blackburn, 'Filling in Space'
- Peter Geach, 'Truth, Duality and Indirect Contexts'
- Martin Hollis, 'Moves and Motives in the Games we Play'
- Frank Jackson, 'Classifying Conditionals'
- Tomis Kapitan, 'Action, Uncertainty and Divine Impotence'
- E. J. Lowe, 'Conditionals, Context and Transitivity'
- Adam Morton, 'Double Conditionals'
- David Papineau, 'Why Supervenience?'
- Philip Pettit, 'Program Explanation: A General Perspective'
- Christopher Ray, 'Paradoxical Tasks'

The Conference papers will be published in an expanded March 1990 issue of ANALYSIS.

For further details and information about registration, write to the Conference Organizer, Professor D. H. Mellor, University of Cambridge, Faculty of Philosophy, Sidgwick Avenue, Cambridge CB3 9DA.

ANALYSIS 50.1 JANUARY 1990

105
An 13

LIAR SYLLOGISMS

By THEODORE M. DRANGE

IT has been shown by Stephen Read that arguments which refer to themselves can yield paradoxes ('Self-reference and Validity', *Synthese* 42, 1979, pp. 265–74). He considers one-premiss arguments, following examples previously discussed by the medieval logician Pseudo-Scotus. Among Read's paradoxical results is that if an argument's only premiss is 'This argument is valid,' then the argument *is indeed valid,* no matter what its conclusion may be. The reason it is valid is that if one assumes *that* premiss to be true, then any conclusion whatever can be derived from it. For if a one-premiss argument is valid and its premiss is true, then its conclusion must also be true. Thus, from the assumption that the premiss 'This argument is valid' is true, the conclusion can be derived. But if the conclusion can be derived from the premiss, then the argument must be valid. So we get the paradoxical result that *any* argument the sole premiss of which says of the argument itself that it is valid is indeed a valid argument!

Carrying the paradox one step further, Read shows how a contradiction can be established if the conclusion of such an argument were in fact false. Consider, for example, the following (call it 'argument (A)'):

(A) (1) Argument (A) is valid.
(2) Hence, snow is not white.

As shown above, (A) is valid, since its conclusion, (A2), can be derived from its premiss, (A1). Thus, premiss (A1) is indeed true! But from our knowledge of snow, we see that the conclusion, (A2), is in fact false. So argument (A) has a true premiss and a false conclusion, which makes it *invalid.* We can thus prove the contradiction that (A) is both valid and invalid.

Another paradox demonstrated by Read is that a contradiction can be derived using any argument with a single true premiss the conclusion of which says that the argument itself is *invalid.* Consider, for example, the following (call it 'argument (B)'):

(B) (1) Snow is white.
(2) Hence, argument (B) is not valid.

If (B) were valid, then it would have a true premiss and a false conclusion, and so would be invalid. Thus, (B) must be invalid. But (B) must also be valid since its conclusion can be derived from its premiss. The derivation is as follows:

> Suppose that (B2) were false. Then (B) would be a valid argument with a false conclusion. But, since (B1) is true, (B) would then be a valid argument with a true premiss and a false conclusion, which is impossible. It follows that (B2) cannot be false. (B)'s conclusion, therefore, must be true.

Since its conclusion can be derived from its premiss, argument (B) is valid. Hence, (B) is both valid and invalid.

Read's paradoxical arguments, however, may well be thought to be suspect. Someone might be inclined to say that the arguments should not be called valid just because devious proofs can be worked out for their conclusions. Certainly neither argument (A) nor argument (B) exemplifies any familiar valid logical form.

What I shall show in the present paper is that self-referential arguments yield paradoxes even when the only criterion of validity is that of logical form and the arguments clearly meet that criterion. I shall look at two-premiss arguments, or syllogisms, that are valid in virtue of having the form modus tollens or the form labelled 'disjunctive syllogism'. Since their paradoxicality is reminiscent of the Paradox of the Liar, I shall call them 'liar syllogisms'.

In what follows, let us take 'sound argument' to mean a valid argument all the premisses of which are true. My first example of a liar syllogism, then, is the following (call it 'argument (C)'):

(C) (1) If argument (C) is valid and both of its premisses are true, then it is a sound argument.
(2) But (C) is not a sound argument.
(3) Hence, (C) is not valid or else its premisses are not both true.

We can see that premiss (C1) is true in virtue of our definition of 'sound'. And (C2) must also be true. If it were false, then the argument would be sound but contain a false premiss, which is a contradiction. So, by this reasoning, premiss (C2) must also be true.

A paradox seems to emerge, then. Argument (C) has the logical form modus tollens, which makes it a valid syllogism, and that makes the first disjunct of the conclusion, (C3), false. But by the reasoning given above, (C)'s premisses are both true. It follows that the second disjunct of (C3) is false. Since both disjuncts of (C)'s conclusion are false, the whole conclusion is false. Argument (C), then, is a valid syllogism with both of its premisses true and its conclusion false. That is indeed paradoxical.

To try to escape the paradox, perhaps we might question the reasoning by which we inferred that premiss (C2) is true. All that was shown was that the assumption that (C2) is false leads to a

P, 6778

contradiction. Instead of concluding that (C2) is true, we might instead label it 'truthvalueless' (which I shall abbreviate 'TVL'), that is, (C2) is neither true nor false but something else. It is reasonable to say that *that* is the label we should attach to any statement if a contradiction can be derived both from the assumption of its truth and from the assumption of its falsity. We saw how the assumption of (C2)'s falsity entails a contradiction. How about the assumption of its truth?

If (C2) were true, then (C) would not be a sound argument. Yet (C) would be a valid syllogism with both of its premisses true, which meets the definition of a 'sound argument'. Thus, if (C2) were true, then (C) would be both sound and not sound. Since a contradiction can be derived both from the assumption of (C2)'s truth and, as shown earlier, from the assumption of its falsity, it follows that (C2) ought to be classified as TVL.

But there is still something paradoxical about argument (C). We derived a contradiction from the assumption that (C2) is true. We also classified (C2) as TVL. It follows from both of these results that (C)'s premisses are not both true. But that entails that the second disjunct of (C3) is true, which in turn makes (C3) itself true. This is paradoxical because, by the definition of 'sound', (C2) and (C3) seem to be saying exactly the *same* thing. How, then, could one be TVL while the other is true?

Another way of formulating the paradox is to make a separate statement outside the argument to the effect that (C) is not a sound argument. Let us label that statement '(OC)' and record the result as follows:

(OC): (C) is not a sound argument.

Since (C2) is TVL, (C) fails to meet the definition of a 'sound argument'. Hence, (OC) is *true*. But here, then, is the paradox. The outside statement (OC) says the very same thing as premiss (C2) within the argument. How can (OC) be true and (C2) be TVL when they say the very same thing? This puzzle is similar to ones encountered in connection with the Paradox of the Liar. Like the Paradox of the Liar, liar syllogisms focus on some statement the truth value of which is paradoxical. And even calling the statement 'TVL' gives rise to puzzles.

The following argument is a still simpler example of a liar syllogism. (Call it 'argument (D)'.)

(D) (1) Argument (D) is not valid or else its premisses are not both true.
(2) But (D) *is* valid.
(3) Hence, (D)'s premisses are not both true.

Argument (D) has the logical form of disjunctive syllogism, so it is indeed a valid argument. That makes premiss (D2) true and the first disjunct of (D1) false. But what about the second disjunct of premiss (D1)? Let us call it '(SD)'. Is (SD) true or false or TVL? If (SD) were true, then the whole disjunction (D1) would be true. So both of (D)'s premisses would be true, which entails that (SD) is false. Thus, the assumption of (SD)'s truth leads to a contradiction. Now suppose that (SD) were false. Then the whole disjunction (D1) would be false. But (D)'s premisses would in that case not both be true, which makes (SD) true. Thus, the assumption of (SD)'s falsity also leads to a contradiction. It seems we are forced to classify (SD) as TVL.

But that leads to a paradox. If (SD) is TVL, what does that make the whole disjunction (D1)? A reasonable rule here is that if one disjunct is TVL and the other is true, then the whole disjunction is itself *true*, but if the other disjunct is false, then the whole disjunction is itself *TVL*. By this rule, the disjunctive premiss (D1) should be classified as TVL. But now we are led to a puzzle similar to the one connected with argument (C). If (D1) is TVL, then (D)'s premisses are not both true. Yet that is precisely what the conclusion (D3) is claiming. Are we to say that (D3) is true? How can (D3) be true when we had classified the very same proposition, in the form of (SD), as TVL?

Another way to formulate the paradox would be to make a separate outside statement about argument (D) to the effect that (D)'s premisses are not both true. Let us label that statement '(OD)' and record the result as follows:

(OD) Argument (D)'s premisses are not both true.

We can definitely say of (OD) that it is *true*, since it follows from the fact that premiss (D1) is not true but TVL. But here, then, is the paradox. (OD) and (SD) say the very same thing. How is it possible for (OD) to be true when (SD) is TVL?

It should be noted that even if we had decided to classify premiss (D1) as *false* (perhaps focussing on its first disjunct), the same paradoxical result would be obtained. (D3) and (OD) would still be true, whereas (SD) would be TVL, yet they are all equivalent, saying the very same thing.

Perhaps one way to respond to liar syllogisms would be to simply dismiss them as 'defective arguments' on the grounds that they are self-referential. Let us call this the 'Dismissal Approach'. It claims that no arguments which refer to themselves can be regarded as properly constructed, so all statements within them, or at least all the ones that refer to the argument, should be automatically classified as 'TVL' and should *not* be seen as being equivalent to outside statements like (OC) and (OD). The appearance that the inside statements and the outside statements say the same thing should be simply taken to be an illusion.

There are several things wrong with this way of dealing with liar syllogisms, but I shall here mention just two. First, it seems counter-intuitive. Premises like (C1) and (D2) seem clearly *true*.

(C1) is true by definition, and (D2) can be seen to be true by simply examining a certain argument's logical form. So, to call such statements 'TVL' seems wrong-headed. Consider, also, the following three arguments:

(E) (1) If this argument (E) is a syllogism, then it contains two premisses.
(2) Well, (E) *is* a syllogism.
(3) Therefore, (E) contains two premisses.

(F) (1) If this argument (F) is sound, then it is valid.
(2) But (F) is not sound.
(3) Therefore, (F) is not valid.

(G) (1) If this argument (G) is sound, then it is valid.
(2) But (G) is not valid.
(3) Therefore, (G) is not sound.

Although these arguments are not liar syllogisms, they are all self-referential arguments and hence to be regarded as defective by the Dismissal Approach. However, our intuitions tell us that (E) is a sound argument, with all of its component statements *true* (not TVL). Argument (F) commits the 'Fallacy of Denying the Antecedent' and so is invalid, hence unsound. But all of its component statements are nevertheless *true*. And argument (G) is valid in virtue of its logical form (modus tollens), which makes its second premise *false*, which in turn makes its conclusion *true*. Thus, (G) is a valid argument with one true premiss, one false premiss, and a true conclusion. It seems reasonable to say these things about arguments (E), (F), and (G). It seems unreasonable and counter-intuitive to just dismiss them as 'defective arguments with all their component statements TVL'.

Another objection to the Dismissal Approach is that not all liar syllogisms are self-referential. Consider the following two arguments (to be called 'argument (H)' and 'argument (J)'):

(H) (1) If argument (J), below, is valid and both of its premisses are true, then it is a sound argument.
(2) But (J) is not a sound argument.
(3) Hence, (J) is not valid or else its premisses are not both true.

(J) (1) Both premisses of argument (H), above, are true or else snow is not white.
(2) But snow *is* white.
(3) Hence, both premisses of argument (H) are true.

Even though neither (H) nor (J) refers to itself, they can both be regarded as liar syllogisms, since they are syllogisms that yield a paradoxical result similar to that of the Paradox of the Liar. Let us see how that is so.

Consider, first of all, premiss (H2). If (H2) were true, then both of argument (H)'s premisses would be true, since the first one, (H1), is true in virtue of our definition of 'sound'. But then the first disjunct of premiss (J1) would be true, which would make (J1) itself true. Since snow is indeed white, (J2) is true. Thus, both premisses of argument (J) would be true. And since (J) has a valid form (disjunctive syllogism), it is a valid argument. (J) would be a valid syllogism with both of its premisses true, which makes it a *sound* argument. It follows that (H2) would be *false*, which contradicts the initial assumption.

Now suppose (H2) were false. Then (J) would be a sound argument, which entails that both of its premisses are true. But the second disjunct of (J1) is false, so the only way for (J1) to be true would be for its first disjunct to be true. But in that case both of argument (H)'s premisses would have to be true, which contradicts our initial assumption that (H2) is false.

Since both the assumption of (H2)'s truth and the assumption of its falsity entail contradictions, we are led to say that (H2) is TVL. But then we get a puzzle similar to the ones for arguments (C) and (D). If (H2) is TVL, then it is not true. It follows that, since argument (H)'s premisses are not both true, the first disjunct of premiss (J1) is false. Since the second disjunct is also false, (J1) itself is false, which means that (J) cannot be a sound argument. But that is precisely what (H2) is saying, which seems to make it *true*, not TVL. Thus, even the claim that (H2) is TVL leads to a contradiction!

Another way of expressing the paradox would be to first formulate an outside statement to the effect that argument (J) is not sound. Let us call that statement '(OJ)' and record it as follows:

(OJ): (J) is not a sound argument.

This statement must be true, for if (J) were a sound argument, then its conclusion would have to be true. But its conclusion says that both premisses of argument (H) are true, which includes (H2), according to which (J) is *not* a sound argument. Thus, if (J) were sound, then it would be not sound. (J) therefore cannot be a sound argument and (OJ) must be true. Here, then, is the paradox. Statement (OJ) and premiss (H2) are saying the very same thing, yet (OJ) is true and (H2) is TVL. How can that be? The Dismissal Approach does not help us here because both (OJ) and (H2) are *outside* statements. That is, they are both outside the argument to which they refer, which is in this case argument (J). The Dismissal Approach applies only to self-referential arguments, and neither (H) nor (J) is a self-referential argument.

We have given two reasons for rejecting the Dismissal Approach as it is usually formulated. Some other solution to the 'Paradox of Liar Syllogisms' is needed. Unfortunately, I do not have any such. Nor do I have a satisfactory solution to the Paradox of the Liar itself. But presumably a solution to one of the paradoxes should be applicable to the other as well.

West Virginia University,
Morgantown, WV 26506, U.S.A.

A VARIATION ON A PARADOX

By ALLEN HAZEN

IN his latest book, Raymond Smullyan continues the ethnography of the Knight-Knave Islands: islands whose inhabitants are all either knights (who speak only truth) or knaves (who speak only falsehood). He observes ([4], p. 18) that no inhabitant of these islands will ever make an assertion of the form 'If I am a knight, then . . .' unless . . . is true. His observation does not depend, however, on knights and knaves being born rather than made: if the islanders sometimes changed their status overnight, the observation would hold with regard to assertions of the form 'If I am a knight (today), then . . .'. Even if they changed by the hour, a parallel observation would hold for their explicitly self-referential assertions of the form 'If what I am right now saying is true, then . . .'. But this form, uttered outside the fortunate isles, is productive of paradox. For consider the sentence

(*) If (*) is true, then Raymond Smullyan is a spy from Betelgeuse.

(The consequent may be varied to taste.) The assumption that a sentence is uttered by an inhabitant of Knight-Knave amounts to an assumption of bivalence, licensing unrestricted use of classical logic. A classical material conditional can only be false if its antecedent is true, so, classically, any conditional whose antecedent asserts the truth of the conditional must be true.

The paradoxicality of (*), however, does not depend on classical logic. Smullyan's own argument for his observation can be adapted, and uses only the rules of Intuitionistic Logic for the conditional. For suppose that

(1) (*) is true.

Then (since we have just assumed that it is true) we may assert

(2) If (*) is true, then Raymond Smullyan is a spy from Betelgeuse.

Modus ponens from (2) and (1) gives us

(3) Raymond Smullyan is a spy from Betelgeuse;

since we have just deduced (3) on the assumption of (1), the Deduction Theorem (a.k.a. Rule of Conditional Introduction) allows us to assert

(4) If (*) is true, then Raymond Smullyan is a spy from Betelgeuse.

This, of course, is simply (*); since we have just asserted it we are committed to

(5) (*) is true,

but then modus ponens allows us to draw the final conclusion, free of any hypothesis, that

(6) Raymond Smullyan is a spy from Betelgeuse.

What we have, then, is a paradox related to the Liar in much the way that Curry's Paradox is related to Russell's: Intuitionistic Logic and the conditional will do the paradoxical job of negation, provided you are willing to consider a new sentence (a new class abstract) for each conclusion. Like the Curry paradox, our variation on Smullyan's observation gives bite to the commonplace that 'Intuitionistic Logic has an impredicative notion of implication', and one way of blocking both paradoxes would be to eliminate this impredicativity by adopting something like Myhill's notion of 'levels of implication'.[1]

University of Melbourne,
Parkville, VIC 3052, Australia

REFERENCES

[1] P. T. Geach, 'On *Insolubilia*', *Analysis* 15 (1954) 71–2.
[2] John Myhill, 'Levels of Implication', in A. R. Anderson, R. B. Marcus and R. M. Martin, eds., *The Logical Enterprise* (New Haven: Yale University Press, 1975) 179–85.
[3] John Myhill, 'Paradoxes', *Synthese* 60 (1984) 129–43.
[4] Raymond Smullyan, *Forever Undecided: a puzzle guide to Gödel* (Oxford: Oxford University Press, 1987).

---

[1] After the present note was submitted for publication, I noticed that the same paradox is presented by Geach in [1].

A DEFENCE OF MODUS TOLLENS

By WALTER SINNOTT-ARMSTRONG, JAMES MOOR, *and* ROBERT FOGELIN

ERNEST ADAMS recently discussed an apparent counterexample to modus tollens.[1] He presents his example first as a set of sentences with no context:

(1) If it rained, it didn't rain hard.
(2) It rained hard.
(3) So, it didn't rain.

Next, he suggests a 'story' or context where one person 'A' asserts (1), and another person 'B' asserts (2). Adams points out that it would be 'implausible' for A to draw the conclusion that (3). This would be equally implausible for B.

Adams realizes that (1)–(3) is not really a counterexample to modus tollens. The reason is that

(4) If it rained hard, it rained.

is 'a logical truth', and (2) plus (4) imply the denial of (1). Thus, B denies what A asserts, so 'it would be irrational for either of them to accept a conclusion that only follows from both of them together.' (pp. 122–3).

However, Adams adds, 'this can't be the whole story' (p. 123). His reason is that in other dialogues it is legitimate to draw conclusions from conflicting assertions by different speakers. His example is:

(5) If Jones took chemistry, he didn't take physics.
(6) He took physics.
(7) So, he didn't take chemistry.

where A asserts (5), and B asserts (6). Adams says that it might be 'rational' for A to conclude (7). The problem is to explain why it is rational to draw conclusions with modus tollens from conflicting assertions in some dialogues but not others.

A simple explanation is that (1) contradicts (2) given (4), but (5) does not contradict (6). Adams dismisses this simple explanation, because it would be 'odd' to assert the conjunction of (5) and (6):

(8) Jones took physics, but if he took chemistry he didn't take it [physics].

so there is also 'a kind of conflict' between (5) and (6) (p. 123).

Instead, Adams offers two complementary explanations. One (pp. 124–5) is that, when A asserts (5), A would also accept the counterfactual:

---

[1] '*Modus Tollens* revisited', *Analysis* 48.3 (1988), pp. 122–8. All references in the text will be to this paper.

(9) If Jones had taken chemistry, he wouldn't have taken physics.

but, when A asserts (1), A would not accept the counterfactual:

(10) If it had rained, it would not have rained hard.

This difference is supposed to explain why (3) cannot be inferred from (1)–(2) in a dialogue, but (7) can be inferred from (5)–(6) in a dialogue.

This counterfactual explanation has been ably criticized by A. J. Dale.[2] Dale gives examples where A would reject the corresponding counterfactual but could still draw the corresponding conclusion.

However, Dale's criticisms do not apply to Adam's other explanation (pp. 123–4). Adams claims that, when A asserts (1), A would not accept the contrapositive of (1):

(11) If it rained hard, it didn't rain.

In contrast, when A asserts (5), A would accept the contrapositive of (5):

(12) If he took physics, he didn't take chemistry.

If modus tollens is acceptable when but only when the speaker accepts the contrapositive of the conditional premiss, this explains why (3) cannot be inferred from (1)–(2) in a dialogue, but (7) can be inferred from (5)–(6) in a dialogue. Adams calls this the tacit modus ponens analysis, since what was modus tollens becomes modus ponens when the contrapositive is substituted for the original conditional.

This tacit modus ponens explanation suggests that (1) cannot be analysed as a material conditional. If (1) is a material conditional, it is equivalent to its contrapositive, (11). Adams claims that this 'would in turn seem to imply that dialogue type *Modus Tollens* inferences should always be rational.' (p. 124). But modus tollens inferences are not rational in dialogues like (1)–(3), so Adams implies that (1) cannot be analysed as material conditional. This makes it important to look at Adams's tacit modus ponens explanation separately.

We will argue that (1) can be analysed as a material conditional, modus tollens is still valid, Adams's tacit modus ponens explanation is inadequate, and there is a better explanation of the apparent failure of modus tollens.

The first step is to understand (1) better by constructing a more concrete context. Suppose A did not see any rain yesterday, but B claims that it did rain. A has some doubt, since A admits that, if it had rained just a little, A might not have seen the rain, although A feels sure that A would have seen rain if it had rained hard. A also wants to avoid argument, since what is important for her purposes is that it did not rain hard. This context makes it natural for A to make a concession to B by asserting (1), possibly with qualifications, as in 'If it rained (at all), then (at least) it didn't rain hard' or '(Even) if it rained (some), it didn't rain hard.' However, B is in a fighting mood, so B responds by asserting (2), and then it does seem wrong for A to conclude that (3).

---

[2] 'Adams on *Modus Tollens*', *Analysis* 49.2 (1989), pp. 93–6.

This context makes it natural to interpret (1) as a material conditional. In asserting (1), A allows two possibilities: either it did not rain at all or it rained but it did not rain hard. This disjunction of possibilities can be symbolized as '$\sim R \vee (R \& \sim H)$', which is equivalent in the propositional calculus to the material conditional '$R \supset \sim H$'. We get the same result when we look at what A denies. When A asserts (1), A denies that it rained hard, so A asserts '$\sim H$'. This implies the material conditional '$R \supset \sim H$' and '$\sim H$' is also implied by '$R \supset \sim H$', given that (4) is analytic, so 'H' entails 'R'. Thus, (1) captures precisely what A intended to allow and to deny when (1) is analysed as a material conditional.

Still, in order to defend the analysis of (1) as a material conditional, we also need to argue that (11), the contrapositive of (1), is true in every case where (1) is. There are two possibilities. First, suppose it really did rain hard. Now it turns out that (1) is false. We can understand why A asserted (1), given the evidence that A had, but that evidence was not conclusive and turned out to be midleading. In this case, (11) is also false, since it did rain hard and it also rained. Next, suppose that it did not rain hard. (1) is then true, and (11) is also true, because its antecedent is false. Thus, neither case provides any reason to reject the equivalence of (1) and (11) or the analysis of (1) as a material conditional.

Why did Adams think otherwise? His main argument is that, even when A asserted (1), 'A would have been very unlikely to have accepted the contrapositive in this case.' (p. 123). However, this observation about what would be *accepted* or *asserted* is compatible with our claims about what is *true*. It is easy to explain why A would assert (1) but would not assert (11) even when both are true. (1) can be used for purposes for which (11) is inappropriate. In our context, the antecedent of (1) is hypothesized as a concession to B who thinks this antecedent is true. This makes it odd to assert (11) as a concession to B, because the consequent of (11) denies what is supposed to be conceded. Moreover, a very common use of conditionals is to warrant inferences to their consequents. But neither A nor B wants to infer the consequent of (11), 'It didn't rain'. Finally, it is more difficult to see A's evidence for (1) as evidence for (11), and this makes it odd to assert (11) instead of (1) on the basis of this evidence. Thus, A's reluctance to accept or assert (11) can be explained without giving up our analysis of (1) as a material conditional.

But maybe Adam's real argument was that (11) cannot ever be *true.* (11) does look like a logical falsehood. However, (11) is asserted only when it is derived by contraposition from (1), and we argued that (1) is a material conditional, so it is natural to analyse (11) also as a material conditional. And then (11) is true by falsity of its antecedent when it did not rain hard. This leaves no reason to deny that (11) is equivalent to (1).

We still need to explain why it seems so strange to conclude (3) from (1)–(2), even if (1) is a material conditional. Our explanation is simple. It is legitimate to draw conclusions from assertions by different speakers only when it is possible that all of the assertions are true together with implicit background assumptions. Otherwise, the argument with its assumptions added cannot be sound, even if it is valid. This makes it illegitimate for A to conclude (3) from (1) and (2), because (1) contradicts (2) given the natural assumption of (4). In contrast, (5) does not contradict (6) given any natural assumption. This explains why it is legitimate to conclude (7) from (5) and (6), even though it is not legitimate to conclude (3) from (1) and (2).

This explanation is supported by the observation that there would be nothing wrong with (1)–(3) without (4). For example, if hard rains were not called 'rains' but 'storms' (just as a hard sprinkle is not a sprinkle but a rain), then (1)–(3) would be satisfactory. So it is with the argument about science courses, since the analogue of (4) would be

(13) If Jones took physics, he took chemistry.

and (13) is not plausible as an assumption. Notice that the implicit assumption need not be analytic in order to make modus tollens paradoxical. Suppose Smith threatened to jump off the World Trade Center. A looks up from the bottom of the World Trade Center, and A does not see Smith jumping, so A says,

(14) If Smith is in danger, (at least) Smith is not jumping off the World Trade Center.

But B does see Smith jumping, so B exclaims

(15) Smith is jumping off the World Trade Center.

In this case, A should not apply modus tollens and conclude

(16) Smith is not in danger.

Here the implicit assumption analogous to (4) is

(17) If Smith is jumping off the World Trade Center, Smith is in danger.

(17) is not analytic, but (14), (15), and (17) are inconsistent and that is enough for modus tollens to produce paradoxical results.

Another strength of our simple explanation is that it applies to arguments other than modus tollens. For instance, suppose that A is not impressed with Robinson as a runner and says,

(18) There is no way that Robinson will run in the race and win.

However, B sees Robinson approaching the finish line first, so B says

(19) Robinson will win the race.

It is not legitimate for A to conclude

(20) Robinson will not run in the race.

even though (20) follows from (18)–(19). This argument has nothing to do with modus tollens or contrapositives, but the explanation is the same as before. The speakers implicitly assume that

(21) If Robinson will win the race, Robinson will run in the race.

(18), (19), and (21) are inconsistent, so the complete argument with (21) as a premiss is not sound. But if (21) is not assumed (e.g. if Robinson can win a political race as a write-in candidate without running), then the argument is not paradoxical. Thus, when the argument is paradoxical, what makes it so is that its premisses and implicit assumptions are jointly inconsistent.

As we said, Adams dismissed this simple explanation, but his only objection was that it would be 'odd' to assert (8), which is the conjunction of (5) and (6), so there is also 'a kind of conflict' between (5) and (6) (p. 123). However, this conflict is not of the right kind. (8) is not contradictory, even though it is odd. We still need to explain why it is odd to assert (8), but the explanation is again easy. As Dale points out,[3] it is the order of (8) that is odd, not its content. It is not odd to assert the conjunction of (5) and (6) in a different order:

(22) If Jones took chemistry, he didn't take physics, but he took physics.

It is common for conjunctions to seem odd in one order but not in the other. Contrast 'He jumped on his horse and rode into the sunset' with 'He rode into the sunset and jumped on his horse.' Similarly, A heard B say that Jones took physics after A already said that, if Jones took chemistry, he didn't take physics. This explains why it is odd to conjoin these assertions in the opposite order. Moreover, conditionals are often used as tickets to infer to their consequents. But (8) denies the conclusion of such an inference before the conditional warrants the inference. There is no point in issuing a ticket to a show that has already been cancelled. Thus, we can explain why (8) is odd without giving up our simple explanation of why it is illegitimate to infer (3) from (1)–(2) given (4).

---

[3] 'Adams on *Modus Tollens*', p. 95.

Another possible objection is that each assertion in (1)–(3) can be justified at the time when it is made. However, that is only because dialogues take time. A asserts (1) before B asserts (2) and later A draws conclusion (3). We need to ask whether or not A agrees with B after B asserts (2). If A then agrees with B, and thus believes that it did rain hard, then A must give up her original belief in (1), given (4). This is no paradox. People change their beliefs during conversation. But after this change, A should not go on to use the rejected belief in (1) to form further beliefs, such as (3). That is why it is paradoxical for A to conclude that (3). The other possibility is that A does not agree with B's assertion that (2). But then A again should not go on to use the rejected belief in (2) to form further beliefs, such as (3). Either way, it is illegitimate for A to use both (1) and (2) in order to conclude that (3), given (4). That explains why the argument seems odd.

Another advantage of our simple explanation is that it saves modus tollens on either of two interpretations. Modus tollens is usually seen as a kind of inference within formal logic. These formal inferences would still be valid, even if (1)–(3) were not valid, since (1)–(3) are not formalized, and part of the issue is how to formalize (1). Another way to see modus tollens is as kind of inference in natural language, such as from English sentences of the forms 'If ..., then ---' and 'not ---' to an English sentence of the form 'not ...'. These inferences would not be valid if (1)–(3) were not valid, but we have argued that (1)–(3) is valid, despite its oddity. Similar inferences in other cases might not be valid, if the conditional premiss is not subject to contraposition and thus is not a material conditional, but that is not shown by Adams's examples.

Even when such inferences are valid, it is not clear when it is rational to believe the conclusion. One might think that, when someone believes English sentences of the forms 'If ..., then ---' and 'not ---', it is rational for that person to form a belief in an English sentence of the form 'not ...'. However, this claim is falsified by (1)–(3), because (1)–(2) are inconsistent given (4), and the rationality of drawing a conclusion in any inference depends on the rationality of believing the premisses. Garbage in, garbage out. In any case, modus tollens remains valid, even when it is not rational to believe its conclusion.

In addition to the validity and rationality of modus tollens, we can also ask about the psychology of modus tollens. Adams claims that a speaker's belief in a conclusion of modus tollens is 'the result of a previous belief in the contrapositive of the conditional premiss (p. 126). His argument is: 'Supposing that to believe something at a time is to attach a high probability to it at that time, for A to come to believe ~C after learning P is for $PROB_1$ (~C) to be high', where '$PROB_1$ (~C)' is the posterior probability of 'Jones did not take chemistry' after learning 'Jones took physics'. Adams equates this with the prior probability of 'if Jones took physics, Jones did not take chemistry' before learning 'Jones took physics',[4] so he concludes that 'prior to learning that Jones did take physics A must have believed or "accepted" the contrapositive conditional "If Jones took physics, he didn't take chemistry", whether he stated it explicitly or not.' (pp. 125–6). Unfortunately, Adams misapplies his supposition about belief. Even if A's beliefs after learning P imply that the prior probability of 'if P then ~C' *is* high, this does not show that A *attached* that high prior probability before A learned P. A did then believe and attach a high probability to 'if C then ~P', and on our analysis this is equivalent to 'if P then ~C', so it would be irrational for A to deny 'if P then ~C'. But this does not show that A must actually form this belief in order to apply modus tollens. Since modus tollens is valid by itself, A does not *have to* derive the contrapositive and then apply modus ponens. Some people sometimes might go through this roundabout reasoning, but there is no reason to think that Adams's tacit modus ponens analysis 'accounts for A's thinking' (p. 126) in all or even most cases.

We conclude that (1) can be analysed as a material conditional, modus tollens is valid, Adams's tacit modus ponens explanation is unnecessary, and our explanation of the apparent failure of modus tollens is simpler and more general. Adams's underlying mistake is to confuse truth and validity with assertion and belief. Adams's confusion is displayed clearly when he declares the lesson of his argument: 'The formalist or abstract semanticist is apt to say that these pragmatic details [where these items of information came from, who supplies them, and so on] are irrelevant to the *validity* of the reasoning, but if I am right they *must* be supplied if it is to be understood and evaluated.' (p. 127; his emphasis). This confuses two issues. The evaluation of an argument does depend on pragmatic details. Pragmatics helps us to identify implicit assumptions as well as the purpose of an argument, and these affect whether an argument is good or bad. But the validity of an argument is a separate issue. Arguments with inconsistent premises are valid, but they should be evaluated negatively, since they are not sound. Thus, both Adams and the formalists are partly right. The formalists are right that validity can be determined apart from pragmatics (in these cases), but pragmatics are still needed for other kinds of understanding and evaluation. The lesson of our argument is that we need both pragmatics and formal logic, but these approaches must be distinguished in order to avoid paradoxes.

---

[4] Adams also equates conditional probabilities with probabilities of conditionals, but we argued against this equation in 'A Defense of Modus Ponens', *Journal of Philosophy* 83, 5 (May 1986), pp. 298–9.

Dartmouth College,
Hanover, NH 03755, U.S.A.

'ACTUALLY'

By ROGER TEICHMANN

MARTIN DAVIES [3, p. 220] has, following Crossley and Humberstone [2], made use of a certain sort of example: a sentence which appears to be untranslatable into a quantified modal language containing only operators for 'possibly' and 'necessarily'. The sentence discussed by Davies is:

> It is possible that everything which is actually red should have been shiny.

This sentence – call it S – is not properly rendered by either of (1) or (2):

(1) $\Diamond(\forall x)(x \text{ is red} \rightarrow x \text{ is shiny})$
(2) $(\forall x)(x \text{ is red} \rightarrow \Diamond x \text{ is shiny})$.

(1) amounts to 'It is possible that all red things be shiny', and (2) amounts to 'Every red thing is such that it might be (might have been) shiny'; and neither of these is equivalent to S.

Sentences like S, Davies argues, indicate the need for a third modal operator, 'actually'. Davies goes on to outline the properties we may regard such an operator as having, and shows some of the uses to which it might be put.

S might well be thought to carry the implication that there are in fact red things. If so, this is more likely to be due to the presence of 'actually' than, for example, the presence of 'everything' (which can carry existential commitment on occasion). For what existential commitment S does carry is surely also carried by the sentence, 'It is possible that nothing which is actually red should have been shiny' – a sentence which poses the same apparent problems for a modal language lacking an 'actuality' operator.

For the moment I will assume that S does imply the existence of red things. Let us, then, consider (S′), a vernacular paraphrase of (S):

> (S′) There are a group of things which are all red, and which alone are red, and it is possible that they should all be shiny.

(S′), note, makes no use of 'actually' or any synonym for 'actually'. How might (S′) be rendered formally? One way, with 'F__' for '__ is red' and 'G__' for '__ is shiny', would be as follows:[1]

(3) $(\exists s)((\forall x)(x \in s \equiv Fx) \ \& \ \Diamond(\forall x)(x \in s \rightarrow Gx))$,

where the initial quantifier 'quantifies over sets. (3) seems to be an adequate rendition of (S) itself, on the reasonable assumption that sets have the same members in different possible worlds (to speak in such terms); though it should be noted that (3) does not imply the existence of red things. But one could quite easily arrange for such an implication by the addition of a clause like '... & $(\exists x)(Fx)$'.

However, one might well think that it would be preferable to do without quantification over sets if at all possible. For one thing, such quantification faces its own difficulties, so that one entangles oneself in those difficulties to the extent that one relies on their source. (Thus, it might turn out that Russell's 'no class' theory of types was the best response to Russell's Paradox – in which case, we could hardly rest content with (3).) Moreover, a philosopher who denies that we need an extra 'actually' operator may well have motives for showing that the operators 'possibly' and 'necessarily' can cope, not only without the help of 'actually', but without the help of the apparatus of possible worlds, or of that of set theory. If the meaning of 'actually' is explained by its proponents in terms of possible worlds (as it generally is), those who have qualms about 'actually' may well be of the sort of nominalist stamp which likewise induces them to have qualms about quantification over sets. (This indeed is my position.)

In fact, another course is open. One way of reading (S′) is as meaning: 'For some positive n, there are exactly n things which are red, and it is possible that they should all be shiny'. In a logicist programme

(4) 'For some positive n, there are exactly n things that are F'

amounts to a quantified sentence whose verifiers are the sentences 'There is exactly one F', 'There are exactly two F's', 'There are exactly three F's', and so on. These sentences are themselves quantificational, corresponding to the formulae in the following series:

$(\exists x)(Fx \ \& \ (\forall y)(Fy \rightarrow y = x))$

$(\exists x)(\exists y)(Fx \ \& \ Fy \ \& \ (\forall z)(Fz \rightarrow z = x \vee z = y) \ \& \ x \neq y)$

$(\exists x)(\exists y)(\exists z)(Fx \ \& \ Fy \ \& \ Fz \ \& \ (\forall w)(Fw \rightarrow w = x \vee w = y \vee w = z) \ \& \ x \neq y \ \& \ x \neq z \ \& \ y \neq z)$

etc.

---

[1] I am in debt here to Timothy Williamson and Peter Smith, both of whom suggested (3) as a rendition of (S′).

A sentence like 'For some positive n, there are exactly n red things' is thus a sort of 'quantification over quantifications': it employs a quantifier the substituends for whose variables are themselves first-order quantifiers (using 'quantifier' to cover such expressions as '$(\exists x)(\exists y)\ (\underline{\quad}(x,y)\underline{\quad})$', as well as '$(\exists x)(\underline{\quad}x\underline{\quad})$'). The details of (various forms of) this logicist approach have been worked out elsewhere (for example, in [1], [4], [5]). A non-objectualist interpretation of the quantifiers is adopted (either a substitutionalist interpretation or one similar); and having defined locutions of the form 'There are n F's', it is possible to go on to define the various forms of mathematical equation – thus '$m+n=k$ can be rendered: 'For all F, for all G, if there are n things that are F and m things that are G, then there are at most k things that are either F or G' – as well as higher-order locutions such as those starting 'For some n ...' or 'For all n ...'.

In order to get 'For some positive n, there are exactly n red things, and it is possible that they should all be shiny', we need only formulate the existential generalization of the formulae in a series that is just like the above series, except in containing extra modal conjuncts: '$\Diamond Gx$', '$\Diamond(Gx \ \& \ Gy)$', '$\Diamond(Gx \ \& \ Gy \ \& \ Gz)$', etc. For present purposes we could just as well read (S′) as meaning: 'For some positive n, there are *at most* n red things, etc.', and since this will make for a rather simpler formal account, I will adopt this reading. But a rendition of 'For some positive n, there are exactly n red things, etc.' is clearly possible also.

Here, then, is the series of formulae from which it is our aim to derive the existential generalization:

$$(\exists x)(Fx \ \& \ (\forall y)(Fy \rightarrow y=x) \ \& \ \Diamond Gx)$$

$$(\exists x)(\exists y)(Fx \ \& \ Fy, \ \& \ (\forall z)(Fz \rightarrow z=x \vee z=y), \ \& \ \Diamond(Gx \ \& \ Gy))$$

$$(\exists x)(\exists y)(\exists z)(Fx \ \& \ Fy \ \& \ Fz, \ \& \ (\forall w)(Fw \rightarrow w=x \vee w=y \vee w=z), \ \& \ \Diamond(Gx \ \& \ Gy \ \& \ Gz)) \text{ etc.}$$

etc.

As with 'For some positive n, there are most n red things' (or 'For some positive n, there are exactly n red things'), the formula in question will employ a quantifier the substituends for whose variables are themselves first-order quantifiers. The inclusion, in the matrices of formulae in the above series, of conjuncts such as '$\Diamond(Gx \ \& \ Gy)$' poses no more problem than does the inclusion of conjuncts such as '$x \neq y$' in the formulae from which we derive the existential generalization, 'For some positive n, there are exactly n F's'.

If it is decided that S carries no existential commitment, this will clearly matter no more than does the fact that measures need to be taken in a logicist programme to allow 'There are no F's' as a possible verifier of 'For some n, there are n F's'. A very simple way of depriving our translation of S of existential commitment would be simply to add: '$\ldots \vee \neg(\exists x)(Fx)$'.

Thus whether we adopt (3), or instead some logicist paraphrase of the sort I have been describing, it appears that sentences like S do not establish the need for a separate 'actuality' operator. The resources of a quantified language containing only 'possibly' and 'necessarily' are adequate to deal with the likes of S – even if not the resources of modal first-order quantification.[2]

Merton College,
Oxford OX1 4JD

REFERENCES

[1] David Bostock, *Logic and Arithmetic,* vol. 1 (Oxford: Oxford University Press, 1974).
[2] J. L. Crossley and I. M. Humberstone, 'The Logic of "Actually"', *Reports on Mathematical Logic* 8 (1977) 11–29.
[3] M. K. Davies, *Meaning, Quantification, Necessity* (London: Routledge & Kegan Paul, 1981).
[4] Dale Gottlieb, *Ontological Economy* (Oxford: Oxford University Press, 1980).
[5] C. J. F. Williams, *What is Identity?* (Oxford: Oxford University Press, forthcoming).

[2] I am grateful to the Editor for his comments on an earlier draft.

ARE SALMON'S 'GUISES' DISGUISED FREGEAN SENSES?

By JOÃO BRANQUINHO

IN a review of *Frege's Puzzle,*[1] Graeme Forbes makes the claim that Salmon's account of belief might be seen, under certain conditions, as a mere notational variant of a neo-Fregean theory; and thus that such an account might be reduced to a neo-Fregean one simply by rewriting it in terms of Fregean terminology. With a view to supporting his claim, Forbes offers an outline of an account of belief which, according to him, would satisfy the following conditions: (i) it could be directly obtained from Salmon's own analysis by means of a certain set of substitutions, which presumably would not affect the essential features of Salmon's view; (ii) it could naturally be described as Fregean, in the sense that it would preserve (at least) the spirit of Frege's doctrines, especially his fundamental intuitions about belief. Of course, the upshot of Forbes's argument is that Salmon's theory would not, at bottom, constitute a genuine alternative to a Fregean semantics for belief ascriptions.

---

[1] Nathan Salmon, *Frege's Puzzle* (Cambridge MA: MIT Press, 1983): reviewed by Graeme Forbes, *Philosophical Review* 96 (1987) 455–8.

In this paper I shall argue to the effect that Forbes's claim is not in general sound. It seems to me that the sort of indirect argument used by Forbes – that of trying to undermine Salmon's theory by showing that it is just a version of a neo-Fregean account – does not provide someone working within a Fregean framework with an adequate strategy to counter Salmon's neo-Russellian views. It would perhaps be better to concentrate a Fregean attack on certain apparently dubious and highly controversial theses and results which are constitutive of Salmon's view, e.g. the counter-intuitive character of a substantial set of consequences which follow from his theory of belief, as well as the associated revisionist stand he is forced to take towards our current patterns of speaking about belief.

Forbes is surely right when he says that Salmon's account implies the possibility of ascribing inconsistent beliefs to some agent. Suppose that our believer, on the basis of his visual encounters with Venus at dawn, thinks that Phosphorus is a star; and that, on the basis of his encounters with Venus at dusk, he thinks that Hesperus is not a star, but a planet. Let us represent Salmon's account by the formula

(S) A believes that p iff (∃x)[A grasps p by means of x & BEL(A, p, x)]

where 'A' and 'p' stand for a believer and a proposition, respectively, and 'x' ranges over ways by means of which A is acquainted with p (*guises*). Using Salmon's characterization of a guise as a binary function f(A,'S'), mapping a believer A and a sentence 'S' onto the way A would take the proposition encoded in 'S' were this proposition presented to him through 'S', the following hold (with respect to our story):

(1) BEL[A, that Hesperus is a star, f(A, 'Phosphorus is a star')]
(2) BEL[A, that Hesperus is not a star, f(A, 'Hesperus is not a star')].

By (S) one obtains from (1) and (2), respectively, the following as true:

(3) A believes that Hesperus is a star
(4) A believes that Hesperus is not a star.

Hence Salmon's theory allows us to attribute contradictory beliefs to A, though it does not allow us to describe him as an illogical or irrational believer: whereas A believes the proposition that Hesperus is a star under the guise f(A, 'Phosphorus is a star'), he believes the contradictory of that proposition under the guise f(A, 'Hesperus is not a star'), and the sentence 'Hesperus is not a star' is obviously not the negation of the sentence 'Phosphorus is a star'. Of course, these observations do not render Salmon's analysis any less counter-intuitive for it has in any case the consequence that A believes that Hesperus is a star *tout court* and also believes that Hesperus is not a star *tout court*; but they somehow mitigate the implausibility of such a consequence.

Forbes's neo-Fregean reconstruction of Salmon's theory consists in subjecting (S) to the following substitutions: (i) on the right-hand side, replace the ternary predicate 'A grasps p by means of x' by 'A thinks of p by means of x', where 'A' stands for a believer and 'p' and 'x' range over states of affairs and propositions, respectively; (ii) still on the right-hand side, replace 'BEL(A,p,x)' by the conjunction 'B(A,x) & p = Ref(x)', where 'A', 'p', 'x' are as above, 'B' stands for the belief-relation, and 'Ref' stands for the Reference function, which in this case assigns to each proposition x some state of affairs as its referent; (iii) substitute 'p is believed by A to obtain' for the left-hand side. The result would be the formula:

(F) p is believed by A to obtain iff (∃x)[A thinks of p by means of x & B(A,x) & p = Ref(x)].

Thus, Forbes's states of affairs replace Salmon's propositions and Forbes's propositions replace Salmon's guises. The underlying idea is that states of affairs are the referents of sentences, or (to be precise) the referents of the propositions expressed by these sentences. It is implicitly assumed that the relation between a proposition x and a certain state of affairs p is the relation which obtains between Fregean sense (the proposition expressed by some sentence 'S') and reference (the state of affairs referred to by 'S'); and this relation is to the effect that sense determines reference, so that we can say that a sentence 'S' refers to a state of affairs p only in a derivative way; primarily, it is the proposition x expressed by 'S' which determines p as its referent. Since his states of affairs coincide with Salmon's propositions – they are indeed conceived as displaying precisely the same composition as Russellian propositions, viz. an object coupled with a property – Forbes contends that Salmon's concept of a guise plays in (S) exactly the same role as the concept of a Fregean sense in (F). Given that a Russellian proposition p is an entity structured in the way described above, every guise under which one can take p is bound to have as constituents a manner of grasping some object *a* and a manner of grasping some property *F*. But such manners of grasping are no other than the Fregean senses which together make up the proposition $\alpha\hat{}\beta$, where 'α' and 'β' stand for modes of presentation of *a* and *F*, respectively (I use '$\hat{}$' for concatenation of senses). Therefore, just as in Forbes's analysis (F) a (Fregean) proposition x is a mode of presentation of some state of affairs p, so in Salmon's analysis (S) a guise x could be viewed as being, in essence, a mode of presentation of some (Russellian) proposition (= state of affairs) p. (I assume that we are only dealing with singular propositions, i.e., those which may be expressed by sentences constructed out of simple unary predicates and simple singular terms.)

P 6778

Although I think that the above considerations show that there must be *some sense* in which Salmon's account could be seen as invoking the same kind of conceptual apparatus as the neo-Fregean theory, I shall argue that Forbes's 'notational variation' claim cannot be accepted; and if I am right, Salmon's guises cannot be assimilated, at least on that basis, to Fregean senses.

If Salmon's (S) is to be regarded as a notational variant of its neo-Fregean counterpart (F), then (F) itself is bound to have the same consequences as (S); in particular, (F) is bound to yield exactly the same verdicts on the truth-values of given belief ascriptions. But either this is not the case, and then Forbes is wrong in claiming that Salmon's account is a notational variant of (F); or it is the case, but then (F) would at once disqualify itself as Fregean (for obvious reasons). I shall now try to show that the first branch of this disjunction holds.

Having in mind A's story one could reason as follows. Let 'x' be ⌜Phosphorus is a star⌝. Then one would clearly obtain the following (for convenience, I abbreviate 'the state of affairs that p' to '/p/'):

(5) $(\exists x)$[A thinks of /Hesperus is a star/ by means of x & B(A,x) & /Hesperus is a star/ = Ref(x)].

And, by (F), we come to the conclusion that

(6) The state of affairs that Hesperus is a star is believed by A to obtain.

On the other hand, by letting 'x' be ⌜Hesperus is not a star⌝ and applying (F), one would also derive the following:

(7) The state of affairs that Hesperus is not a star is believed by A to obtain.

So the conjunction of (6) and (7) implies at least that A believes of inconsistent states of affairs that they both obtain. But is there any sense in which this could amount to ascribing contradictory beliefs to A?

In order to answer this question we have to determine whether or not (6) and (7) might receive, respectively, the readings:

(8) A believes that the state of affairs that Hesperus is a star obtains

(9) A believes that the state of affairs that Hesperus is not a star obtains.

If so, I would argue on Fregean grounds that (8) and (9), and hence (6) and (7), are respectively the same as (3) and (4), in which case (F) would indeed entail the possibility of our ascribing inconsistent beliefs to A. Taking (8) as an example, the occurrence of the singular term 'the state of affairs that Hesperus is a star' there must have as its reference, not its customary reference (viz. a certain state of affairs, e.g. the reference of the sentence 'Phosphorous is a star'), but that which is its customary sense; and, using Forbes's assumptions, this sense is a concept of that state of affairs, viz. the Fregean proposition ⌜Hesperus is a star⌝ (which might be expressed by the sentence 'Hesperus is a star' occuring by itself).[2] Thus, the Fregean regimentation for (8) is

(10) B[A, ⌜Hesperus is a star⌝^⌜obtains⌝],

which may be presumably abbreviated to

(11) B[A, ⌜Hesperus is a star⌝],

which in turn is the Fregean regimentation for (3). (A similar argument could be applied to (9).)

However, there is a strong reason to think that Forbes would not allow (6) and (7) to have, respectively, the readings (8) and (9). The reason is that e.g. (6) would rather receive (in Forbes's terminology) the 'external form' reading

(6*) The state of affairs that Hesperus is a star is something A believes to obtain

where the occurrence of the singular term 'the state of affairs that Hesperus is a star' is an external occurrence and has as its reference the customary reference of the term (the state of affairs itself), being thus open to replacement *salva veritate* by any co-referential singular term. By contrast, in (8) the occurrence of the same term is internal and (8) is itself an internal form. Now the logical relations which are seen to hold between external forms and the corresponding internal forms is that the former ascriptions can always be inferred from the latter ones, while the converse deduction is not valid in Forbes's theory; that is, whereas a belief-ascription of the general form

(*) *a* is something A believes to be *F*

is entailed by a belief-ascription of the form

(**) A believes that *a* is *F*,

the converse inference is not allowed. It follows that (6), read as (6*) which has the form (*), does not entail (8), which has the form (**). Thus, (6) cannot have the reading (8), and the same would apply to (7) and (9).

Our general conclusion is that, under such circumstances, Forbes's analysis (F) does not yield the same consequences, concerning the truth-values of belief-ascriptions such as (3) and (4), as

---

[2] Actually it could be that the description 'the state of affairs that p' might express a slightly different (perhaps more intimate) way of thinking of a state of affairs than that which is expressed by the sentence 'p'; but I ignore the complication here.

Salmon's analysis (F); for instance, whereas (S) implies that (3) is true (in our story), (F) would entail the falsity of (3). This is sufficient to establish the result that, *pace* Forbes, Salmon's theory is not a notational variant of a neo-Fregean account (as represented in (F)). Nevertheless, and independently of translation considerations, Forbes's reconstruction has the merit of enabling us to see Salmon's theory, from a Fregean point of view, as misinterpreting belief-ascriptions by treating states of affairs as the objects of belief.[3]

New College,
Oxford OX1 3BN

---

[3] I am very grateful to Timothy Williamson and to Graeme Forbes for their helpful remarks. I also thank the anonymous referee. This paper was written when I held a Calouste Gulbenkian Foundation Fellowship.

HONDERICH ON THE INDISPENSABILITY OF THE MENTAL

By GREGORY McCULLOCH

IN his recent book, *A Theory of Determinism*, Ted Honderich presents a series of criticisms of various types of mind-brain identity theory (see [4], chapter 2). In particular, he attacks the mildest version of the identity theory, namely the bare claim of token-by-token identity between mental events and neural events (as they occur in humans and physiologically similar creatures: I shall follow Honderich in ignoring this qualification). The attack is foreshadowed in an earlier paper on Anomalous Monism, and is elucidated in a subsequent debate with Peter Smith (see [1], [2], [3], [6], [7]). In this article I am concerned with a weakness in Honderich's argument which does not emerge from that debate.

The argument pivots on what Honderich calls the *axiom of mental indispensability* (hereafter 'the Axiom'), whose initial formulation is as follows:

> earlier mental events are essential to any full explanations of certain later mental events and also actions. ... There is no full explanation of my reaching for an olive which does not mention my wanting one, no full explanation of my wanting one which does not mention my seeing them on the table (p. 91; all unattributed page references are to [4]).

He maintains that the Axiom

> derives from a fact on which there is agreement by almost all philosophical parties. ... The fact ... must be regarded as ... pre-theoretical or

> pre-philosophical, something against which theory and philosophy are to be tested rather than a proposition derived from theory or philosophy (p. 91).

And he continues that it

> is not in itself the proposition that earlier mental events *cause* later mental events and actions, although such causation will indeed make for indispensability (p. 91).

He then argues that the bare claim of token-token identity between neural and mental events (hereafter Token-identity) 'does not satisfy the requirement of mental indispensability' (p. 96). This, given the previous remark, means that the argument is supposed to show that Token-identity fails to secure the claim

(1) Earlier mental events cause certain later mental events and also actions.

Such a result would be very surprising and significant, had it been established. But we shall see that (1) admits of at least two interpretations, only one of which is an embarrassment to the proponent of Token-identity. Further, we shall see that although there is indeed an undeniable fact about the explanation of action which may appear to force the stronger, embarrassing reading of (1) on to the proponent of Token-identity, this pressure can be resisted.

A key step in Honderich's argument involves his view that 'The most natural and explicit answer to the question of what caused something ... is simply a property or relation of an ordinary thing' (p. 15). Although we 'often refer to causes as *events*', 'it is as natural' also to say that 'facts', or 'ordinary things or indeed *stuff*, such as the snow' are causes (p. 14). But such locutions, says Honderich, including the event-locution, are derivative ways of speaking 'in accord with the rooted practice of taking the whole for the part' (p. 16). His view is that *strictly speaking*, only properties or relations (or sets thereof) are causes, and that therefore, in particular,

(2) Causal talk of mental events is strictly speaking talk of properties of those events (see p. 110).

Now if we grant (1) and (2) we should be hard pressed to deny the further claim

(3) Some properties of earlier mental events cause later mental events and also actions.

So suppose decision D is brought about by cogitation C. Then (1) requires that C caused D. Moreover, if Token-identity is true there will be some neural event N_1 identical with C and a further neural event N_2 identical with D. Furthermore, it would not be very plausible to deny that N_1 caused N_2. On the contrary, it is with such causal claims that proponents of Token-identity are likely to seek to deliver what (1) requires: it is an integral part of identity theories that, as Honderich puts it 'neural events ... have certain causal roles' (p. 93) which causal roles are supposed to underpin those attributed to mental events. But given the claims that $C=N_1$, $D=N_2$, and N_1 caused N_2, one might have thought that the desired result that C caused D is available in the circumstances, even if it does not follow immediately by two applications of Leibniz's Law. For it seems obvious that if we are allowed to say that C *qua* neural event caused D, then we can hardly be prevented from saying that C caused D, and thereby satisfying (1) and hence the Axiom. If a pear *qua* thing with weight causes a pointer to move, then the pear causes the pointer to move. If I *qua* owner of a loud voice keep the audience awake, then I keep the audience awake. In other words, proponents of Token-identity can cheerfully accept (3) and then propose

(4) The neural properties of earlier mental events cause later mental events and also actions,

and will, surely, consider themselves thereby to have satisfied (1).

Yet Honderich insists there is an error here. Token-identity, he says, is the claim that 'there is a single entity, which entity has both a mental and a neural [property]' (p. 94). However, he goes on, 'it is essential to see that [this claim] does not give us the conclusion that all the ... properties of the single entity have the same standing' (p. 95). What he means is this. In appealing to the causal roles of neural events such as N_1 and N_2 in order to underpin the causal roles of the likes of C and D, proponents of Token-identity may well appeal merely to the causal efficacy of the *neural* properties of the events in question: certainly no more than this is implied by the bare claim of token-by-token identity. Hence no causal role is secured for the *mental* properties of those events, so that Token-identity, at least in the context of (2), itself yields a rather muted causal claim about the mental.[1] And Honderich immediately infers from this a fatal *explanatory* weakness in Token-identity:

> That the neural part of the entity has a certain causal role does not

---

[1] Here I disagree with Cynthia and Graham Macdonald (see [5] pp. 36–40). A detail I have suppressed is that Honderich regards *individual* properties to be causes (thus *my voices's* loudness, rather than the multiply instantiable loudness, is what keeps the audience awake). The Macdonalds interpret this as the claim that it is the instance of a property (i.e. the individual which has it), rather than the property itself, which is the cause: and then they unsurprisingly find only confusion in Honderich's claim that merely displaying a token mental event as a cause in virtue of its being a neural event makes a muted causal claim about the mental event. For if the event (*qua* instance of whatever neural property it has) is thereby shown to be a cause, and to be an instance is all a cause, strictly speaking, can be — then what more does Token-identity need to achieve? But Honderich distinguishes individual properties from their instances — he claims they stand in some sort of part-whole relationship (see e.g. [4] p. 16) — so the Macdonalds simply miss his point.

give us the upshot that the mental part is *an ineliminable part of any explanation of a later thing* (p. 99, my emphasis).

But what is his argument here? The immediate thought is that Honderich supposes that (1) and (2) somehow lead not merely to (3), but also to

(5) The mental properties of earlier mental events cause later mental events and also actions.

And (5) is certainly not captured by (4) in the absence of further claims, such as identity claims about neural and mental properties. But if (1) and (2) yield (5), and (4) *did* secure (1), then only rejection of (2) could stand between the proponents of Token-identity and (5). So, short of turning attention on (2) with a view to denying it, proponents of Token-identity must accept that (4) fails, despite appearances, to accommodate (1).

But we can continue to allow (2) to pass unexamined, for the mistake here is evident. If we allow that talk of event-cause is strictly speaking talk of property-cause then we should allow that talk of mental event-cause is talk of property-cause. But it seems sophistical to insist further that talk of mental event-cause is talk of *mental* property-cause. Suppose that talk of certain wholes is strictly speaking talk of square parts. Does it follow that talk of the round wholes is strictly speaking talk of round square parts? Hardly. Furthermore, even if there were an entailment of the required sort here, that would merely show that the widespread belief in the Axiom to which Honderich appeals is not, given (2) and despite first appearances, adequately expressed by (1), and is better captured by (3). For in invoking (5) Honderich clearly intends (1) to carry the import that all adequate or full causal explanations of certain mental events and also actions must be *mentalistic*; and belief in this is simply not so widely unquestioned as his strategy requires. On the contrary, there is a widespread and growing belief that advances in neuroscience and the like could make possible full and adequate explanations of certain mental events and also actions which will definitely not be mentalistic. And there is no apparent reason why this cannot be conjoined with belief in (3).

Is Honderich guilty of such a simple error? Perhaps not, for there is a rather more subtle possibility of confusion here. He notes that if we 'absolutely exclude the mental realistically conceived from the explanation of the bodily sequence, so too do we exclude an intentional action' (p. 213). That is, it is constitutive of something's being an action that we can see it as the appropriate kind of upshot of prior deliberations etc. *mentalistically conceived.* This is indeed a fundamental fact about our Intentional conceptual scheme, essential to the distinction between action and mere bodily occurrence, and in this sense it is certainly subject to 'agreement by almost all philosophical parties' that there is no proper or full explanation of an action which is not mentalistic. That is, we cannot deny that

(6) An adequate or full explanation of an action as conceived mentalistically must itself be mentalistic.

But then (6) might seem to combine with (2) to yield the substance of (5), and thus deliver the kind of constraint on causal explanations of action which Honderich requires. Given (2), an explanation of an action is strictly speaking to do with properties. But given (6), these properties themselves must be mental — for how else is the explanation to be mentalistic? Hence, if (1) is *correctly understood* — that is, understood in the context of (2) and (6) — then it can indeed be expressed as (5), and the case against Token-identity seems to be made out after all unless (2) can be resisted.

However, we still do not need to examine (2), for there remains a flaw in the argument. Quite simply, why should the expected future explainers of the doings of epistemic engines and the like conceive these doings as actions? Of course, no-one has the faintest idea how these doings will be conceived, but is there an argument which shows that they will simply have to be conceived as actions *on pain of explanatory or predictive weakness*? I doubt that such an argument could be forthcoming. It will certainly not do merely to point out that abandoning mentalism would be contrary to our present explanatory practice, or that it would be *hard to get by*, on a practical level, without resorting to mentalism. What is needed, if this causal-explanatory sort of defence of mentalism is to be successfully mounted, is an *a priori* refutation of the possibility of any descriptively and explanatorily complete account of human behaviour that is not mentalistic. And that, to say the least, is a very tall order indeed, and is certainly not provided in Honderich's rather perfunctory dismissal of eliminative materialism (pp. 103–4).

In effect, then, I suggest that Honderich's argument equivocates between the two glosses of the Axiom given as (3) and (5) above, the equivocation being camouflaged somewhat by the presence of (6). And whilst both (3) and (6) are indeed widely and firmly believed, the argument needs (5) to go through. But (5) is far from widely and firmly believed, once belief in it is distinguished from belief in (6): worse, (5) is not even particularly believable in the absence of the appropriate kind of causal-explanatory argument against hard-nosed eliminativists. So those who believe that Token-identity *itself* secures the causal efficacy of the mental can rest easy for a little longer.[2]

University of Nottingham,
University Park, Nottingham NG7 2RD

---

[2] I am grateful to the Editor for helpful comments on an earlier version.

REFERENCES

[1] Ted Honderich, 'The Argument for Anomalous Monism', *Analysis* 42 (1982) 59–64.
[2] ——, 'Anomalous Monism: Reply to Smith', *Analysis* 43 (1983) 146–9.
[3] ——, 'Smith and the Champion of Mauve', *Analysis* 44 (1984) 86–9.
[4] ——, *A Theory of Determinism* (Oxford: Oxford University Press, 1988).
[5] Cynthia Macdonald and Graham Macdonald, 'Mental Causes and the Explanation of Behaviour', in *Mind, Causation and Action*, edited by L. Stevenson, R. Squires and J. Haldane (Oxford: Oxford University Press, 1986) 35–48.
[6] Peter Smith, 'Bad News for Anomalous Monism?', *Analysis* 42 (1982) 220–4.
[7] ——, 'Anomalous Monism and Epiphenomenalism: Reply to Honderich', *Analysis* 44 (1984) 83–6.

VAN INWAGEN ON THE 'OBVIOUSNESS' OF LIBERTARIAN MORAL RESPONSIBILITY

By SAUL SMILANSKY

VAN INWAGEN'S recent *An Essay on Free Will* [7] has received much critical attention. This attention has, however, focussed on the purported proofs of incompatibilism offered by van Inwagen and on related matters dealing with the compatibilism-incompatibilism issue (e.g. [1], [2], [3], [6], [8]), to the neglect of another crucial difficulty in van Inwagen's argumentation: the claim that there is a conclusive case for the direct obviousness of the existence of libertarian moral responsibility (i.e. libertarian-free-will-assuming moral responsibility), despite the lack of a case for such direct obviousness as to the existence of libertarian free will. Van Inwagen's arguments at this point seem to me to be very weak. But if this is indeed so, then the whole structure of argumentation of the book is imperilled.

The 'Argument From Introspection' for the existence of libertarian free will is not commonly presented today, since even libertarians usually agree that it is not convincing. It seems to be widely recognized that the *lack* of causes of one's actions is just not the *sort* of thing that could be known by introspection. There is no good reason to think that people must always be aware of the causes of their behaviour. Hypnotized persons provide dramatic illustration of this, as they are commonly not aware of the causes of their behaviour, causes obvious to outsiders. However, there is no need to argue further for this point here, since van Inwagen admits that the 'Argument From Introspection' fails with regard to libertarian free will. He imagines a mechanism in our brain, putting us at the liberty of another, of which we may not be aware ([7], p. 109f.). In fact, van Inwagen even says at one point that the idea that we can find out whether we have libertarian free will by introspection 'seems just obviously wrong ...' (p. 204).

However, van Inwagen seems to think that the reality of libertarian moral responsibility *can* be proved in a way that he himself admits fails in the case of libertarian free will: the existence of libertarian moral responsibility is, in some unclarified way, immediately obvious, while this is not so with libertarian free will. As he puts it, 'surely we cannot doubt the reality of moral responsibility?' (p. 206). Van Inwagen states repeatedly that (in the libertarian sense) we all believe that people are sometimes morally responsible (pp. 208–9), and claims that to deny the existence of such moral responsibility would be 'absurd' (p. 223).

But why is the case for libertarian moral responsibility so outstandingly strong? Van Inwagen does not discuss the issue of moral responsibility with any thoroughness (p. 206f.). We all just know, it is claimed, that we are sometimes morally responsible in the libertarian sense. This of course would seem to contradict what many philosophers have claimed. Van Inwagen's main argument seems to be that everyone *actually* agrees that we have libertarian moral responsibility, and that those who deny libertarian moral responsibility cannot be taken seriously, because they speak and behave in ways which contradict this (p. 207). The main difference between the case here and the case with libertarian free will, according to van Inwagen, is that 'The philosopher who denies free will continually contradicts himself because his nonverbal behaviour continually manifests a belief in free will. But, I would suggest, the philosopher who denies moral responsibility speaks *words* that contradict his theory, words like "That was a shoddy thing to do"' (p. 207).

These arguments are not very convincing, however. Even if people often speak and behave as if others were morally responsible in the strict (libertarian-free-will assuming) sense, this would not show that they cannot believe that people are not morally responsible in this sense, only that they cannot internalize and actualize all their convictions. Thomas Nagel illustrates this: 'I can no more help holding myself and others responsible in ordinary life than I can help feeling that my actions originate with me. But this is just another way in which, from some distance outside, I seem to myself to be trapped' ([4], p. 123). People can come to believe that what they do (and perhaps often cannot refrain from doing), the way in which they sometimes react and speak, is unjustified.

One simple point is this: society has such a great incentive to inculcate in us all the relevant reactions and beliefs that it is not required that such beliefs be founded for them to exist. Nor is it surprising that even hard determinists cannot always act according to their beliefs. As van Inwagen would surely agree, the taking of responsibility and the attribution of it to others in a way in which people might be said to justly deserve gain or loss for their actions, is very important for sustaining a stable social structure. And such social needs (and, historically, religious needs) for the possibility of talking about moral responsibility are greatly furthered if moral responsibility and desert are thought to exist in the deep, libertarian, sense. Even if there is no libertarian moral responsibility, it is only to be expected that the corresponding practices persist. And this point is strengthened once we take into account the common human 'phenomenology of freedom', which is conducive to the impression that we are free and morally responsible in the libertarian sense. For in situations of choice, and in particular hard choice, people commonly feel that they possess libertarian free will ([5], pp. 69–70). And this deeply ingrained feature is naturally seen by them to enable libertarian moral responsibility.

The widespread belief in the existence of libertarian moral responsibility thus has a sufficient explanation both in the common experience of choice (which is interpreted as the experience of having libertarian free will), and in the social ethos. And the combined effect of these elements naturally makes it very difficult to actualize in any thorough way the belief that libertarian free will (and thus libertarian moral responsibility) do not exist. And so, the words and deeds of determinists, even if we accept van Inwagen's characterization of them, are not a proof that (all) determinists do not believe in the lack of libertarian moral responsibility. And this disproves his claim for the (presumably universal) obviousness of the existence of libertarian moral responsibility.

Van Inwagen curiously seems to speak as though there are two disconnected sets of beliefs: belief in the existence of libertarian free will, and belief in the existence of libertarian moral responsibility. However, surely the right thing to say is that — since people generally believe in the existence of libertarian free will — it is not surprising that they believe in the reality of libertarian moral responsibility, i.e. in people being sometimes accountable for their libertarian free actions. And so, if the belief in libertarian free will has no evidential value, the closely related belief in libertarian moral responsibility cannot possess such value.

It is not clear whether there is anything to make of van Inwagen's purported distinction between (1) the determinist's *actions* (in deliberating) contradicting their belief in the existence of libertarian free will; and (2) their *words* contradicting their beliefs (and words) in the case of belief in libertarian moral responsibility. If determinists are indeed at fault as to libertarian free will, they surely express this in words as well, outside of contexts of holding people morally responsible. And so, were this sort of problem — the determinist's belief in determinism being sometimes contradicted by his words — an indication that determinists do not really believe in libertarian moral responsibility, it would apply just as well to the belief in libertarian free will. Van Inwagen would be able to discredit the determinists on the level of libertarian free will, without any need to consider moral responsibility. (Notice, however, that even according to van Inwagen this would still in no way establish the existence of libertarian free will.) And in fact, since it is impossible to make sense of libertarian beliefs existing with regard to moral responsibility but not with regard to free will, it seems that any argument about the sincerity of determinists would have to apply to both elements of the libertarian belief. But as we saw, such arguments are unsuccessful.

To summarize. The existence of libertarian moral responsibility is far from being obvious: many people have doubted this and still doubt it. Since libertarian moral responsibility depends on the at best problematic notion of libertarian free will, it is highly implausible to see the existence of libertarian moral responsibility as obvious; and this is even more implausible if the existence of libertarian moral responsibility is thought to be obvious independently of the case for libertarian free will (as many libertarians would agree). The beliefs of (at least some of) the people who doubt the existence of libertarian moral responsibility cannot be discredited even if they often behave in ways which seem to imply belief in libertarian moral responsibility: the naturalness of belief in libertarian free will and the reinforcement of the belief by the social ethos can explain this. The widespread belief in the libertarian picture, and the strength of the influence of this picture even on those who see reason to doubt its validity, are easily accounted for, even if this picture is unfounded. Both with the belief in libertarian free will and with the closely related belief in libertarian moral responsibility, the common belief in them lacks significant evidential value.

What is the significance of the implausibility of van Inwagen's case for the obviousness of libertarian moral responsibility? It would seem that if we take out this edifice, the whole structure of *An Essay on Free Will* collapses. But in order to see why this is so we must briefly review the (fairly lopsided) structure of this book.

In practice the book is focused on proving compatibilism wrong. However, this is not how van Inwagen seems to perceive the nature of his argumentation — rather, the thrust of the argument is perceived as supporting the existence of libertarian free will. Van Inwagen briefly discusses the question of the coherence of libertarian free will in his chapter 4, but he does not claim to have any significant new thing to say about it. The main point of his discussion is that he does indeed find the libertarian free will conception 'puzzling' (p. 149). His general argument is, however, that in view of his having proved incompatibilism (and since moral responsibility requires free will), alternative positions (e.g. that libertarian free will does not exist) are even more puzzling, or 'inconceivable' (p. 149). But if van Inwagen does not claim to know how its existence is possible, why is the lack of (libertarian) free will 'inconceivable'? The answer is that this would mean that we were not morally responsible in the libertarian sense: 'If we do know that moral responsibility exists, then we should have no doubt about whether we have good reason to believe we have free will. It is this and *only this*, I think, that provides us with a reason for believing in free will' (p. 209, my emphasis). Since we have seen that van Inwagen's case for the obviousness of the existence of libertarian moral responsibility cannot hold, his case for the existence of libertarian free will goes as well.

Even if the core of the work that van Inwagen does in the book, the proof of incompatibilism, is deemed successful, the most that he achieves is to make a case for hard determinism over compatibilism. The case for libertarian free will has to be made directly, by attempting to make sense of this conception. And the tenability of belief in libertarian moral responsibility would depend on the case for belief in libertarian free will – and not vice versa.[1]

Magdalen College,
Oxford OX1 4AU

REFERENCES

[1] John Martin Fisher, 'Van Inwagen on Free Will', *Philosophical Quarterly* 36 (1986) 252–60.
[2] John Martin Fisher, 'Freedom and Miracles', *Nous* 22 (1988) 235–52.
[3] Terence Horgan, 'Compatibilism and the Consequence Argument', *Philosophical Studies* 47 (1985) 339–56.
[4] Thomas Nagel, *The View From Nowhere* (Oxford: Oxford University Press, 1986).
[5] Galen Strawson, *Freedom and Belief* (Oxford: Oxford University Press, 1986).
[6] Richard Swinburne, *Responsibility and Atonement* (Oxford: Oxford University Press, 1989).
[7] Peter van Inwagen, *An Essay on Free Will* (Oxford: Oxford University Press, 1983).
[8] Gary Watson, 'Free Action and Free Will', *Mind* 96 (1987) 155–7.

---

[1] I am grateful to Galen Strawson and to the Editor for comments on drafts of this paper.

HOW TO *MIND* ONE'S *ETHICS*: A REPLY TO VAN INWAGEN

By V. ALAN WHITE

A FEW years ago in this journal Peter van Inwagen attempted to prove that two traditional arguments for compatibilism cannot both be sound ('On Two Arguments for Compatibilism', ANALYSIS 45.3, June 1985, pp. 161–3). Citing several familiar compatibilist texts which employ the two arguments, van Inwagen condenses various expressions of the arguments into his own succinct restatement of them:

THE *ETHICS* ARGUMENT

> Analysis shows that statements of ability are disguised conditionals. More exactly, the correct analysis of 'X could have done A' is 'If X had decided (chosen, willed ...) to do A, X would have done A'. Therefore, having acted freely – having been able to act otherwise than one in fact did – is compatible with determinism (with the causal determinism of one's acts).

THE *MIND* ARGUMENT

> If one's acts were undetermined, they would be 'bolts from the blue'; they would no more be *free* acts than they would if they had been caused by the manipulations of one's nervous system by a freakish demon. Therefore, free action is not merely *compatible* with determinism; it *entails* determinism (pp. 161–2).

(The arguments are so called by van Inwagen because of certain coincidences of publication, nothing more.)

In order to prove that at best only one of these arguments can be sound, van Inwagen first argues that the *Ethics* argument allows that free action is compatible with the indeterminism of *deliberation* by an agent (pp. 162–3). With this result in hand, van Inwagen then draws his intended conclusion:

> If the *Mind* argument is sound, then free will entails determinism. If free will entails determinism, then free will is not compatible with indeterminism. Therefore, if either [the *Mind* or *Ethics*] argument is sound, then the other is unsound (p. 163).

Van Inwagen's proof is unassailable, but only if the *Mind* and *Ethics* arguments are about the very same sense of 'free will'. (Or, at van Inwagen's pleasure, 'free action'. Note that van Inwagen's characterizations of the arguments utilize the latter term, while his final proof utilizes the former. Clearly they are interchangeable terms as far as he is concerned.) However, it apparently never occurs to van Inwagen that these two arguments might be used to make quite distinct points about free will.

I will not rehearse van Inwagen's argument that an *Ethics* analysis is compatible with the indeterminism of deliberation because it is unnecessary (and perhaps unavailing) to resist that conclusion. After all, *Ethics*-style arguments were introduced by compatibilists in the first place because they show that an agent's ability to act otherwise is fully compatible with *any* intelligible account of deliberation, deterministic or not.

What is the *Mind* argument about? A casual rereading of van Inwagen's version supports the impression that it, like its companion argument, is about free action. And, it is true that the argument might be generously interpreted to refer loosely to any and all aspects of human action from deliberation to bodily movement. However, to suppose that the compatibilist's interest in the argument extends to proving the causal determinism of indisputably physiological processes of behavior such as neural transmission or muscular contraction seems rather silly. The *Mind* argument therefore is clearly and specifically directed at shunning indeterministic accounts of deliberation by holding that the vicissitudes of indeterministic mental processes rob agents of control of and responsibility for their actions.

Thus, despite van Inwagen's assumption to the contrary, there are at least *prima facie* reasons for believing that the *Ethics* argument concerns the relation between the free actions of agents and their decisions or choices, independently of any specific account of their prior deliberations, whereas the *Mind* argument insists on the need for a deterministic account of those deliberations. Furthermore, if these arguments are so understood, then not only are they compatible with each other, but they may be combined to yield a composite picture of free will as the free action of agents whose deliberations are causally determined.

In fact, since van Inwagen's purpose is to charge compatibilists with inconsistency, I submit that, by failing to recognize how these arguments are reconcilable in the way I have suggested, his reading of at least one compatibilist's use of these arguments is erroneous. The case in point involves his citation of R. E. Hobart (Dickinson Miller) as an advocate of the *Mind* argument in the classic 'Free Will as Involving Determination and Inconceivable Without It' (*Mind* 43, January 1934, pp. 1–27). Hobart does indeed use the *Mind* argument, but only to show that a concept of causally determined mental processes is a necessary part of a proper conception of free will. Van Inwagen fails to see that Hobart employs the *Ethics* argument as well in order to show that conditional freedom of action is the *other* part of that conception. Hobart was not merely trying to be long-winded in titling his essay as he did: he was signaling that in spelling out what free will is, he would *Mind* his *Ethics*, so to speak.[1]

University of Wisconsin Center,
705 Viebahn Street,
Manitowoc, WI 54220-6699, U.S.A.

---

[1] I thank the Editor and anonymous referees for helpful remarks on earlier drafts of this note.

PATERNALISM DEFINED

By DAVID ARCHARD

PATERNALISM is essentially the usurpation of one person's choice of their own good by another person. In this article I provide a more formal and more precise definition of paternalism which, I believe, improves on others that have been offered. I will introduce a number of examples to illustrate and defend my own definition.

My definition is as follows:

P behaves paternalistically towards Q iff:
(1) P aims to bring it about that with respect to some state(s) of affairs which concerns Q's good Q's choice or opportunity to choose is denied or diminished;
(2) P's belief that this behaviour promotes Q's good is the main reason for P's behaviour;
(3) P discounts Q's belief that P's behaviour does not promote Q's good.

Each of these conditions needs separately to be explained and defended. (1) may be seen as controversial. For most, paternalism is defined as an interference with another's freedom for the purpose of promoting that other's good; and for some the interference must be coercive. However, Bernard Gert and Charles Culver have shown that paternalism need not coerce nor interfere with freedom of action ('Paternalistic Behaviour', *Philosophy & Public Affairs* 6, 1976, p. 46). They instance a doctor who knows of a casualty patient's religious objections to transfusion, but proceeds with one when the patient is unconscious. The doctor acts paternalistically yet does not coerce, or seek to change the behaviour of the patient.

An unconscious patient cannot act. To that extent both her behaviour and her freedom to behave in certain ways cannot be interfered with. Nevertheless, with respect to a state of affairs affecting her welfare, namely her treatment in hospital, her choice not to have a transfusion is denied. (This choice may not have been made explicitly — the patient may have been unconscious on admission — but it is one which may be inferred from what is known of her views. It is the choice she *would* have expressed had she been conscious.) Although she is not in a state to do what she wants to do, she does remain somebody to whom something she does not want done can be done. This example is thus covered by my condition (1).

Gert and Culver argue for a condition with plays the same role as my (1) but which states that P behaves paternalistically towards Q if P violates some moral rule with regard to Q. However, Gerald Dworkin has an example to show that this condition is not necessary ('Paternalism: Some Second Thoughts', in *Paternalism*, edited by Rolf Sartorius, Minneapolis: University of Minnesota Press, 1983, p. 107). A husband is under no moral obligation to inform his suicidal wife of the whereabouts of *his* sleeping pills. He acts paternalistically when he fails to do so, but violates no moral rule by his failure. This example is also covered by my condition (1). The husband seeks to diminish the wife's opportunity of choosing suicide by denying her access to the pills.

The example also, and importantly, demonstrates that P can deny or diminish a choice both by the performance of some action *and* by a failure to perform an action which need not be obligatory but which P would normally have been expected to perform. In the normal course of events the husband would have told his wife where his sleeping pills were to be found.

P may aim but fail, due to the actions of others, to bring it about that Q's choices are denied or diminished. P's behaviour remains paternalist. Again, P may sincerely claim not to aim to bring about a diminution of Q's choices; she seeks only to bring about a state of affairs whose probable and foreseeable but not intended consequence is such a diminution of Q's choices. The doctor withholding a diagnosis, and thereby limiting the kinds of choices her patient can make about her future welfare, may see herself as *merely* saving the patient from the extreme distress of knowing the worst. However, if the consequence of an intentional behaviour may reasonably be foreseen and is the diminution of another's choice, then I believe that the behaviour should be seen as satisfying condition (1) of my definition. Moreover, to the extent that condition (3) of my definition is also true it is unlikely that someone could not intend, let alone foresee the choice-diminishing consequences of their behaviour.

For P's behaviour to be paternalist its intended or foreseeable effect must be to decrease and not simply to change Q's choices. Consider the example of 'the secret patron'. P believes that Q's having or doing X would be for Q's good. X may be two weeks at a health farm, six sessions with a psychoanalyst, or a season ticket to the opera. P knows that Q does not currently believe X to be for Q's good. However, P makes X freely available to Q intending that Q will choose X. Since P knows that Q would be disposed not to choose X if Q knew P to be its donor, P lies to Q about the source of X.

The secret patron's behaviour is paternalistic by conditions (2) and (3). She hopes to promote Q's good, but knows that Q does not believe X is for her good. It would be paternalistic by Gert's and Culver's condition, for P breaks the moral rule of honesty with regard to Q. Nevertheless, P's behaviour *increases* Q's choices who can now choose X at no cost to herself where previously she could not. It would sound perverse to describe P's behaviour with regard to Q as paternalist. (It may be less perverse to say of P that she has a paternalist reason for behaving as she does. I will return to the distinction between a paternalist reason for a behaviour and a paternalist behaviour when discussing condition (2).)

The reason is surely that the effect of P's action is to *add* her choice of what is for Q's good to those choices Q already has, whereas paternalism is about the *supplanting* of someone's choice of her good by that of the paternalist. It is interesting to consider how the example might be viewed if it were the case that P knew Q to be a hopeless sucker when it came to free offers, unable to resist accepting them; or if P's offer was such that Q, notwithstanding her views about its value to her, would feel compelled to accept, or could not reasonably refuse. In these instances, P seems knowingly to be making a coercive offer and thus seeking to bring it about that Q's choices are diminished or denied.[1]

Condition (2) of my definition says that P's belief that her behaviour promotes Q's good must be the *main* reason for her behaviour. Some definitions suggest that a paternalist behaviour is for reasons *exclusively* to do with the other's good. I think this is a mistake. Consider a wife who tries to stop her husband from drinking. She believes that sobriety and abstention are for his own good. But she is also well aware that he becomes violent towards her when drunk. Teetotal, the husband would not abuse her; the improved quality of life he might also enjoy would in turn make for a better marriage and a happier life for his partner.

In general terms, P may believe that a behaviour promotes the good of both P and Q, and P's good may or may not depend on that of Q. Indeed we can imagine a range of cases in which the balance between concern for one's own good and for that of the other varies from almost exclusive preoccupation with oneself to the exact opposite. Should we call paternalist any behaviour done because of some interest, however small, in promoting the other's good? I think the answer lies in distinguishing between a paternalist behaviour and a paternalist reason for a behaviour. The latter is a belief that a behaviour promotes the good of some other, together with a belief that this first belief is not shared by the other, and of course a desire to promote the good of the other. It then seems reasonable to say that a paternalist behaviour is one which satisfies condition (1), condition (3) with a qualification yet to be made clear, and whose main reason is paternalist.

---

[1] To B. Lynn I owe the nice parallel example of the testator who imposes conditions upon Q's inheritance – stipulating, for example, that she not marry a certain person, take a certain job, etc. – which are enforced by the will's trustees. The testator's motives are paternalistic since he believes that satisfaction of the will's conditions are for Q's good, whatever Q might think. Yet Q's choices are increased, not diminished by the existence of the will – save perhaps beyond that point when the inheritance is so large as to make of the will's terms a coercive offer.

This is the force of condition (2). The sense of 'main' is left imprecise but I hope that it is intuitively clear enough.

Condition (3) of my definition expresses the fact that the paternalist ignores the views of the other with regard to her good. Four points need to be made. First, the paternalist may well be mistaken about the views of the other, and it may in fact be the case that the other shares the paternalist's view about what is for her good. That is, P falsely believes that Q believes P's behaviour does not promote Q's good, since Q believes P's behaviour does promote Q's good. This can be so even if P's behaviour satisfies condition (1).

Imagine that Q lets everyone think that she is a habitual but willing and happy user of hard drugs. In fact she is not; she has never even used hard drugs. Q finds herself in a situation where she must choose between taking drugs and revealing her previous deception. Acting as Q's friend and ignorant of the deception, P resolves to release Q from what P views as Q's addiction and destroys the drugs. With respect to the taking of the drugs, P's action has curtailed Q's ability to choose. P acts (we may assume here exclusively) to promote Q's good, and discounting what she believes to be Q's belief, namely that the destruction of the drugs does not serve Q's good. Is P's behaviour paternalist?

There are reasons in favour of both a positive and a negative answer, but I am not sure we need to decide one way or the other. If we do want to reach a decision then we can make use of the distinction between a paternalist reason and a paternalist behaviour. The former is as defined before; crucial to this definition is how the paternalist sees things. However, we may now define a paternalist behaviour as one done mainly for a paternalist reason whose belief concerning the other's belief is true. Should we think that a behaviour is paternalist whatever the status of this belief then we can drop this qualification.

P's destruction of Q's drugs is an example of what might be called misguided paternalism. The mirror image of this, which might be called misguided beneficience, can quickly be dispensed with. Imagine that P believes that her behaviour promotes Q's good, and falsely believes that Q also believes P's behaviour promotes Q's good. Now P's behaviour may well satisfy conditions (1) and (2) of my definition, at least in so far as P intentionally brings about a situation such that Q's choices are diminished. However, since P thinks that Q's shares P's judgement of Q's good, condition (3) is not satisfied, and the behaviour does not count as paternalist.

The second point to be made about condition (3) is that it covers cases both in which P does and in which P does not have a belief about Q's belief. P may discount Q's belief in the sense of overriding Q's own choice of what is best for Q. P is aware of what Q thinks, but judges Q to be mistaken. On the other hand, P may simply not be interested in what Q believes, and form no view of Q's own attitude. P discounts Q's belief in the sense of choosing for Q whatever Q might think. It remains true I think that a behaviour is only paternalist if it would be reasonable for P to conclude, were she to think about it, that her view of what promotes Q's good does not coincide with that of Q herself.

The third point to be made about condition (3) concerns P's warrant for her belief about Q's belief. P should take account not only of Q's currently stated or inferable views but, where appropriate, of any previous statements. This is especially true of prior commitments: an individual may attempt to bind herself to a course of action by asking others to ignore or even override any later expressions of will which run counter to that course of action. Q may, for example, declare that she is giving up smoking and ask that any future requests for a cigarette be ignored. If P restrains Q from lighting a cigarette, the latter's protests may be taken by P as expressions of an aberrant will, the real beliefs of Q with respect to smoking and her well-being having been voiced in the prior declaration of intent to abstain.

Of course we might simply say that in so precommitting herself Q had given up her right to make future choices of a certain kind. To that extent, P does not deny Q a choice that is Q's to exercise in relation to her own good. Thus condition (1) of the definition is not satisfied, and P's behaviour fails to count as paternalistic. Such an interpretation raises crucial and complex questions concerning the legitimacy of contracting to limit one's future freedom which cannot be discussed here.

A fourth point is that (3) may be thought ambiguous in respect of what is meant by 'Q's belief'. Imagine that P dashes from Q's hands (a shouted warning proving ineffective or likely to be so) a cup of Paraquat, judging (correctly) that Q thinks it contains lemonade. Is P's behaviour paternalistic? It is by conditions (1) and (2), but what of (3)? Inasmuch as Q believes herself to be about to drink lemonade, she believes that P's stopping her does not promote her good. And P certainly discounts *this* belief. However, P can claim to be taking account of a belief Q has to the effect that being protected from unwitting ingestions of poison does promote Q's good. Nevertheless, this belief is not one Q entertains with respect to P's behaviour at the time of P's behaviour. It is not the belief that explains Q's choice of action; the belief which does explain Q's decision is the one which P discounts, namely that lemonade is about to be drunk and being stopped from drinking from this cup is not to Q's advantage. It is because P discounts this belief that he behaves paternalistically; it is because this belief is palpably false, and will subsequently be discounted by Q herself, that P can claim to be justified in his paternalism.

It remains true – and my third and fourth points above only broach the issue – that it can be hard unambiguously to specify what an individual's beliefs about her own good are, especially once the possibility of inconsistency in an individual's set of beliefs is granted. Consider someone who chooses to smoke and who may nevertheless have reasonably imputed to them the belief that the ingestion of noxious and probably lethal substances is not in one's interest. How would a snatcher of cigarettes from people's lips compare with the dasher of the cup of Paraquat? Such questions have to do with the justifiability of particular paternalistic behaviours, and are thus distinct from definitional issues. From the perspective of my proposed definition the following is true: whatever other beliefs an individual may, or may be assumed to have, she is the object of a paternalistic behaviour if, at the time of the behaviour and in relation to that behaviour, her belief about what promotes her own good is discounted.

The question of justification having been mentioned, it should be made clear that my definition of paternalism is morally neutral, covering cases of both justified and unjustified behaviour. I think this definitional approach is licit. Some might think that paternalism is by its nature an illegitimate form of behaviour. However, I have argued that paternalistic behaviours should not be defined as those which violate some moral rule with regard to their objects. I have captured in my definition what is essential to paternalism, namely that it is the usurpation of one person's choice of their own good by another. Inasmuch as we are inclined to respect the choices of others then we will of course tend to view all paternalism as unjustified and to be morally suspicious of any paternalist.

Nevertheless, it is clear that we should not always respect the choices human beings make – the decision to drink something not knowing it to be poisonous is an obvious example. Indeed, paternalism is normally thought to be justified in those circumstances where the paternalist has good reasons to believe that an individual's current choice is not the one she would make if certain conditions were fulfilled. On the standard liberal account, these conditions are usually specified in terms of knowledge of the consequences of the chosen course of action and a capacity to make a free and reasonable choice. The liberal seeks to secure, through justified paternalism, the decision which the other person would *really* want to make.

Now what an individual's beliefs and/desires *really* are with respect to their own good can be contentious; what an individual *would* choose if certain conditions were fulfilled will depend upon how these conditions are specified. It would, I believe, be a mistake to build into a definition of paternalism a question-begging presumption of what shall count as an acceptable individual choice of one's own good. Let us agree that paternalism usurps choice, and move on afterwards to discuss what kinds of choice we should honour. Put another way, we have duties to promote the well-being of others, but we also have a duty — at least on the standard anti-paternalist account — to respect the choices of others. Any paternalistic behaviour ignores that duty. It is a further question whether such disregard is justified because a particular choice is of a certain kind, or because the duty to care for others takes precedence over the duty to respect their own choices of their welfare.[2]

University of Ulster at Jordanstown,
Co. Antrim BT37 0QB, Northern Ireland.

---

[2] I would like to thank the Editor for his very helpful comments on an earlier draft of this article.

IS THE BEST REALLY NECESSARY?

By E. R. KRAEMER

STEPHEN GROVER ('Why Only the Best is Good Enough', ANALYSIS 48.4, October 1988, p. 224) defends the Leibnizian view that God would not have created any world at all if there were no best of all possible worlds by arguing that 'the easiest way' to defend God's being perfectly good is to claim that there is no possible world better than the world which God chooses to create. I am not convinced.

Let us suppose that there were no best of all possible worlds but only an infinite series of marginally improving possible worlds. An omnipotent God, Grover claims, could create any of them, but a 'weak god' could create many but not all members of the series. Were this not the best possible world, Grover argues, there could be a marginally better possible world that some weak god could create. This, he says, would 'impugn His [perfect] goodness', as there could be a being other than God who could act better than God. Suppose we grant that the existence of such a weak god would impugn God's goodness. What is not clear is why appealing to claim (A), that this is the best of all possible worlds, is any easier a way to defend God's goodness than the alternative of relying on claim (B), that there is no possible weak god who could create a world better than the world which God chooses to create. Were there no such possible being, God's goodness could not be impugned since nothing could act better than God.

But, it might be objected, what is the difference between (A) and (B)? If there is no possible weak god who could create a world better than the world which God chooses to create, then, the objection continues, this must surely also be the best possible world. To show that there does seem to be a real difference between (A) and (B) it should be noted that *only* if we were also to assume such controversial claims as (1) that no world can exist without being created by a god, and (2) that God must create the very best world that He can (even if there are a series of marginally improving worlds which *only* He could create), will (B) yield the same result as (A). So, until such claims as (1) and (2) can be established as correct, we should remain confident that (A) and (B) are indeed importantly different.

As God's perfect goodness may be protected using either (A) or (B), it remains to be considered whether either of these two strategies is any easier to maintain than the other. I submit that strategy (B) is no harder to defend than strategy (A), since similar objections can be raised against both strategies. For example, it might be claimed that (B) must be false as we can well imagine that there could be a weak god who could create a world better than the one which God chooses to create. But, it might also be urged that (A) must be false since we can surely conceive of a world much better than this one. Not finding any way of defending (A) over (B), I suggest that, for God's goodness to remain inviolate, this need not be the best of all possible worlds, so long as there is no other possible being who could create a world better than this one.[1]

University of Wisconsin — La Crosse,
La Crosse, WI 54601, U.S.A.

---

[1] I am indebted to comments from the Editor.

INTUITIONISM VERSUS DEGREES OF TRUTH

By STEPHEN P. SCHWARTZ

IN 'Degrees of Truth versus Intuitionism' (ANALYSIS 49, January 1989, 31–2) George Rea compares Putnam's intuitionist proposal[1] for dealing with sorites paradoxes with the degrees of truth approach. Rea has little sympathy for the intuitionist proposal. Among other things Rea claims that 'the intuitionist proposal, in effect, renders intuitionist logic useless' (p. 31). This is because the intuitionist can always escape acceptance of a superficially good argument by denying the conjunction of the premisses without denying any particular premiss (or even holding that at least one premiss is false, since $\neg(P \& Q)$ does not entail $\neg P \vee \neg Q$ in intuitionist logic). 'Even though all the premisses seem undeniable, and the argument seems valid, the conclusion need not be accepted on the grounds that the conjunction of the premisses might nevertheless be false' (p. 32). He says that intuitionism 'opens up a free for all escape route' (p. 32). Rea thinks this gives intuitionism some sort of merely *ad hoc* advantage over other logics. But contrary to what Rea claims, intuitionism allows for no special 'free for all escape route'. Whereas I may agree with Rea that the intuitionist approach does not ultimately offer a solution to the sorites paradox,[2] I think that he has failed to appreciate the distinct advantages that intuitionism has over classical logic and degrees of truth as a logic of vague terms.

---

[1] See my 'Intuitonism and Sorites', *Analysis* 47 (October 1987) 179–83, for a discussion of this proposal.

The 'free for all escape route' Rea refers to is nothing peculiar to the intuitionalist proposal. The fact is that if an argument is valid and has a false conclusion, then we have no choice but to deny the conjunction of the premisses. This is the case whether we are using intuitionist logic or classical logic (or degrees of truth): Validity of argument plus falsity of conclusion means that the conjunction of the premisses cannot be true. Thus if we believe of an argument that it is valid and that its conclusion is false, we are rationally committed to disbelieving the conjunction of the premisses. This is not an escape route, it is just logic.

It may seem that the intuitionist is getting away with something when he says that the conjunction of the premisses is false but is unwilling to deny any of the conjuncts. Classical logic, in effect, allows the same thing however. Although we are rationally committed to disbelieving the conjunction of the premisses (because the argument is valid and has a false conclusion), this does not mean that we are rationally committed to disbelieving some particular premiss. It is certainly possible, classically, to reject a conjunction without rejecting any particular conjunct. Classically, if I disbelieve a conjunction C, then I must believe $(\exists x)(x$ is a conjunct of C & x is false$)$ but it need not be the case that $(\exists x)(x$ is a conjunct of C & I believe that x is false$)$. In other words if a classicist rejects a conjunction, he must hold that at least one conjunct is false but he need have no beliefs about which conjunct is false. So the classicist and the intuitionist are about on a par with respect to 'free for all escape routes'. The intuitionist can say the conjunction is false but refuse to deny any particular conjunct. The classicist as well can say that the conjunction is false but refuse to deny any particular conjunct.

---

[2] I admit I have been a bit overly optimistic in speaking of 'an intuitionist solution' to the sorites paradox in 'Intuitionism and Sorites'. A lot depends on what conditions one places on 'solution to a paradox'. The problem with the intuitionist 'solution' is that it lacks independence (as do the other 'solutions' to the sorites paradox). The intuitionist offers no *independent* reason (no reason other than its leading to paradox) for rejecting the conjunction of the premisses. See Stephen P. Schwartz and William Throop, 'Intutionism and Vagueness' (forthcoming in *Erkenntnis*).

This is precisely the situation facing us in the case of the sorites argument. In the case of the expanded version of the sorites[3] we are faced with an argument that is valid and has a false conclusion (that a man with 10^8 hairs on his head is bald). We have no choice but to deny the conjunction of the premisses. In the case of the sorites, however, this commits the classicist (but not the intuitionist) to precision because in denying the conjunction of the sorties premisses the classicist must accept not just that *at least* one premiss is false but that *exactly* one premiss is false.[4] This means the classicist is committed to an exact, to the hair, dividing line between bald and non-bald.

The advantage of the intuitionist approach is that the intuitionist is not, just by denying the conjunction of the premisses, committed to a sharp dividing line. Since the intuitionist rests falsity on refutation, he can deny a conjunction without being committed to holding that at least one of the conjuncts is false (i.e. refutable). Obviously, we can be in the position to refute a conjunction without having any reason to suppose that any of the conjuncts is refutable. For example, $P \mathbin{\&} \neg P$ is refutable even though neither P nor $\neg P$ may be. So the intuitionist is able to avoid the unwanted, at least for the logic of vague terms, commitment to one conjunct being false. Thus the intuitionist avoids the commitment to dividing line precision.

That the intuitionist is not committed to dividing line precision is a distinct advantage of intuitionist logic as a logic of vague terms. To impose a precise dividing line on vague terms as the classicist must (and as the degrees of truth theorist must, as I shall shortly argue) is essentially to give up on vagueness. The intuitionist, on the other hand, can take vagueness at face value. By 'take vagueness at face value' I mean can treat vague terms as vague as opposed to treating them as containing somewhere a hidden precision.

---

[3] See my 'Intuitionism and Sorites' (p. 181), for this version.

[4] See my 'Intuitionism and Sorites' (p.182) for the explanation of why this is so. Basically the idea is that the premisses (all but the first) are conditionals. They are so related in the baldness sorites that if one is false (true antecedent, false consequent), then the true antecedent makes each of the previous consequents true and the false consequent makes each of the subsequent antecedents false. The false conditional would be of the form 'If a man with n hairs is bald, then a man with $n+1$ hairs is bald'. But a man with fewer hairs than a bald man is bald and a man with more hairs than a non-bald man is non-bald.

Like the intuitionist proposal, the degrees of truth approach has the advantage of avoiding a precise dividing line between bald and non-bald but the degrees of truth approach is committed to precision nevertheless and thus, like classical logic, is unable to take vagueness at face value.

In *The Metaphysics of Modality* (Oxford: Oxford University Press, 1985) Graeme Forbes, whom Rea cites with approval, supposes that each vague sentence would be assigned a real number in the interval between 0 and 1 reflecting its truth value (p. 170). Contrary to vagueness this is the most refined and unbelievable precision.

Other degrees of truth theorists, recognizing this problem, suppose that a vague but borderline sentence would have an indeterminate truth value somewhere in the interval between 0 and 1.[5] Even if we allow indeterminate truth values, degrees of truth is nevertheless still committed to a form of dividing line precision. Let '$B(n)$' be 'A man with exactly n hairs on his head is bald'. Consider the set $\{B(10^8), B(10^8-1), \ldots, B(0)\}$. Presumably $Val(B(10^8)) = 0$ (in other words, it is definitely false that a man with 10^8 hairs is bald) whereas for some $n < 10^8$, $Val(B(n)) > 0$. Presumably there will be many n for which the truth value of $B(n)$ will be somewhere between 0 and 1. This is the whole point of the degrees of truth approach. Borderline cases will get intermediate truth values between completely false and totally true. But since the set $\{B(10^8), B(10^8-1), \ldots, B(0)\}$ is finite and the values are well-ordered, there must be a first n such that $Val(B(n)) > 0$. In other words there will be a precise, to the hair, dividing line between definitely non-bald and borderline non-bald. This means that there is an absolute precision implied by the degrees of truth approach that is inconsistent with the vagueness of 'bald'. A precise and unknown dividing line between definitely non-bald and borderline is just as contrary to the vagueness of 'bald', and just as unbelievable, as would be a precise dividing line between non-bald and bald. The same kind of argument could presumably be repeated for any vague term.

It seems then that the degrees of truth approach in its treatment of vagueness is committed to dividing line precision just as is classical logic. This is a crucial failing, since to impose dividing line

---

[5] But then it is hard to understand their definitions of the logical functions. Rea gives the standard degrees of truth clause for the conditional as follows:

$$\begin{aligned} Val(A \rightarrow B) &= 1 - (Val(A) - Val(B)) \text{ if } Val(A) > Val(B) \\ &= 1 \text{ otherwise} \end{aligned}$$

The problem is that the arithmetic function of subtraction is not defined for indeterminate quantities. In arithmetic what are subtracted are real numbers. If '$-$' is being used to represent some other function than the arithmetic function of subtraction, this should be made clear.

precision on vague terms is to give up on taking vagueness seriously. The advantage of intuitionism over degrees of truth as well as classical logic as a logic of vague terms is that intuitionism is not committed to such precision.[6]

Ithaca College,
Ithaca, NY 14850, U.S.A.

---

[6] I would like to thank my colleague Frederik Kaufman and my wife Diane Driscoll Schwartz for helpful suggestions in writing this paper.

A DOXASTIC PARADOX

By JON PEREZ LARAUDOGOITIA

SUPPOSE that person A is required to respond to a number of statements in a test: to each statment he is asked to answer sincerely saying whether he himself believes it or not. Let's imagine a situation in which for many of the statements that A has to give a response, it happens that he believes them, although he allows that they might be false, or does not believe them, but with the suspicion that they might be true. Thus, he sensibly suspects that, at least once, he has made a mistake, either believing something false or else not believing something true. Finally, suppose that the final statement of the test affirms: 'Something that you have affirmed to believe in the test is untrue or [inclusive disjunction] there is something true which you have affirmed not to believe'. How would A respond to this? From what has been said above, it is obvious that he would affirm that he believes it. Let's call the last statement of the test H, and ask ourselves if A made a mistake in his belief. If H is false, then all the statements believed by A would be true and those he didn't believe would be false. As A believes H, H must be true, a contradiction. Therefore, H must be true. This sounds sensible enough and A would like to be shown, at some time, which statements that he affirmed to believe are false or which statements he affirmed not to believe are really true. Normally, this shouldn't cause any trouble, but now let's suppose that it turns out that A has made no mistakes (leaving H aside for a moment); that is to say, that everything he earlier affirms to believe in the test is true and what he affirms not to believe is false. Obviously, our earlier demonstration that H cannot be false still applies, and thus H is true. However, this cannot be the case because now, in fact, it is false that A believes a test statement which is false, and it is also false that he does not believe a test statement which is true (in particular H is true and A believes it so). In other words, the hypothesis that A believes H leads to contradiction and, by *reductio ad absurdum,* we can conclude that it is false that A believes H.

A, like most people, is reasonably modest and sincerely affirms that he believes H. In most cases a belief of this sort is correct. However, we have just seen that there are situations in which A does not believe H in spite of saying so. This is pragmatically absurd: how can the contingency that A has made no mistake earlier in the test debar him from coherently believing H? Thus, a variant of the preface paradox[1] has brought us to a still stranger doxastic paradox.

Consider now another person, B. And let's suppose B to be a vain person convinced of his presumptive infallibility. Suppose too that B has answered to our test in the same way as A except that B affirmed not to believe H. Then, we find ourselves in a similar situation to the truth-teller paradox (cf. [2]), for H can be either true or false without contradiction; it is impossible to decide between the alternatives, although the former leads immediately to one of Moore's paradoxes (cf. [3]).

Universidad del País Vasco,
Vitoria, Spain

REFERENCES

[1] Bernard Linsky, 'Factives, Blindspots and some Paradoxes', *Analysis* 46 (1986) 10–15.
[2] C. Mortenson and G. Priest, 'The Truth Teller Paradox', *Logique et Analyse* 24 (1981) 381–8.
[3] J. N. Williams, 'Moore's Paradox: One or Two?', *Analysis* 39 (1979), 141–2.

---

[1] Even though this paradox originated with Makinson, I refer here to the interesting version given by Linsky [1] following the lines of an argument by MacIntosh.

Vol. 50 No. 2 (New Series No. 226) March 199

ANALYSIS

Edited by
PETER SMITH

Moves and motives in the games we play	MARTIN HOLL
Filling in space	SIMON BLACKBUR
Why supervenience?	DAVID PAPINEA
Paradoxical tasks	CHRISTOPHER RA
Double conditionals	ADAM MORTO
Conditionals, context and transitivity	E. J. LOW
Why is belief involuntary?	JONATHAN BENNET
Program explanation: a general perspective	FRANK JACKSON *and* PHILIP PETTI
Why did Psammenitus not pity his son?	AARON BEN-ZEE
Action, uncertainty and divine impotence	TOMIS KAPITA
Classifying conditionals	FRANK JACKSO

BASIL BLACKWELL · OXFORD AND CAMBRIDGE, MA

ISSN 0003-2638

ANALYSIS 50

Newnham College, Cambridge Conference Programme

Friday 30 March

5 p.m.	Inaugural Address: P. T. Geach, 'Truth, Duality and Indirect Contexts'
8.30 p.m.	Martin Hollis, 'Moves and Motives in the Games we Play'

Saturday 31 March

10 a.m.	Simon Blackburn, 'Filling in Space'
11.30 a.m.	(i) David Papineau, 'Why Supervenience?' (ii) Christopher Ray, 'Paradoxical Tasks'
3.30 p.m.	Adam Morton, 'Double Conditionals'
5 p.m.	E. J. Lowe, 'Conditionals, Context and Transitivity'
8.30 p.m.	Jonathan Bennett, 'Why is Belief Involuntary?'

Sunday 1 April

10 a.m.	Philip Pettit, 'Program Explanation: A General Perspective'
11.30 a.m.	(i) Aaron Ben-Zeev, 'Why did Psammenitus not Pity his Son?' (ii) Tomis Kapitan, 'Action, Uncertainty and Divine Impotence'
2.30 p.m.	Frank Jackson, 'Classifying Conditionals'

ANALYSIS 50.2 MARCH 1990

MOVES AND MOTIVES IN THE GAMES WE PLAY

By Martin Hollis

WHEN Jack and Jill went up the hill to fetch a pail of water, they were playing a one-shot, two-person game where both emerged as losers. Had they been the cool, clearheaded, rational agents of game theory, they might both have done better. Then again they might not. Game theory is a wondrously austere abstraction from our own tumbledown world of laughter and forgetting, where the heart has its reasons and ignorant armies clash by night. Perhaps something crucial is lost in the abstraction, so that Jack and Jill are an allegory for the just deserts of rational maximizers. In Gary Becker's view 'a person decides to marry when the utility expected from marriage exceeds that expected from remaining single or for an additional search for a more suitable mate'.[1] If such a person finds a like-minded mate, one is inclined to say, it will serve them both right. But pause before deciding that Becker has a foolish approach to a trip to the well. Game theory embodies a philosophically enticing claim to portray and advise our enlightened selves and, if social life is not that sort of game, we must go on to say what other sort of game it is.

Social action is action which, in Max Weber's useful phrase, 'takes account of the behaviour of others and is thereby oriented in its course'. To study social action, game theory takes a further leaf from Weber's book and sets up an ideal-type world where all agents are fully rational. In the crucial limiting case each rational agent has fully ordered preferences, complete information and perfect skill in computing. Preferences are so well ordered for possible outcomes that the agent always knows how he ranks any pair of them. Information is so complete that the agent always has a subjective probability distribution for all uncertainties which he faces; and everyone knows what anyone knows, since each can reproduce the premisses and reasoning of each other agent. Each has the similar aim of (broadly speaking) maximizing utility; and each knows also that each has this aim. It is, in short, common knowledge that all agents are rational in the sense just given, a situation marked in the ideal-type case by calling them 'hyperrational'.

Philosophers often find this notion of rationality, or hyperrationality, quite *outré* when it is sprung on them. But, basically, it is the familiar Humean notion abstracted to a limiting case and so not really *outré* at all. I therefore ask for patience while I deploy this idea that social life consists of games played by rational individuals with common knowledge. Having done so, I shall then grant that there is something very odd about an ideal-type world which lacks what I shall term the *grit, dust* and *glue* of our own. But, I shall argue, complaints on this score must be finally directed at the familiar Humean notion. The question will then be how thoroughly to overhaul it.

---

[1] See his *The Economic Approach to Human Behaviour* (Chicago: Chicago University Press, 1976) pp. 3–14, for the full story.

To lend human interest to austere abstraction, some games have acquired names. There is, for instance, 'Battle of the Sexes', a name to delight James Thurber or Sigmund Freud. In this game Jack and Jill are contemplating an evening at either a bullfight or a concert. Each would prefer to have the other's company, although Jack would rather that both went to the bullfight and Jill favours the concert. Their choices being made separately, there are four possible outcomes, which each ranks as the pay-offs indicate in *game 1.*

		Jack	
		Bullfight	Concert
Jill	Bullfight	2,3	1,0
	Concert	0,1	3,2

game 1

Thus, if both choose the bullfight, Jill will gain 2 utils and Jack 3 utils, these being the pleasant units of some notional felicific calculus. (For the moment suspend queries about the quantities and what they might represent; and do not presume that interpersonal comparisons of utility are involved.) Both players gain from choices which work out cooperatively but do not have the same order for these outcomes.

The game has two Nash-equilibria or outcomes where neither player can make a better choice, given the choice by the other (setting aside 'mixed strategies' like tossing a coin — a complication which I shall ignore). If Jack knew that Jill was heading for the concert, he would rationally do the same; if Jill knew he was on his way to the bullfight, she would rationally go too. But neither knows until choices have been made, and the Thurberesque battle is on.

Does either have a reason to pick a particular option? For a one-shot game without communication the answer is No. Any reason for Jack to do one thing would be a reason for Jill to do the other, with an inferior outcome for both. That sounds terribly austere. What if he loves her more than she loves him, a fact known to both? Will that not make the concert salient? The answer is again No. Any such difference is already included in the pay-offs. Jack's '2' in the bottom right box is made up of his pleasure in the music, in her company and in relishing her enjoyment. Jill's '2' in the top left box need not be made up in analogous proportions. Game theory stands aloof from the details of their separate motivations.

Yet that sounds too bland. So let me check by juggling with the pay-offs a little, as in *game 2*.

		Jack	
		Bullfight	Concert
Jill	Bullfight	3,4	0,0
	Concert	1,1	5,3

game 2

Here, if we may lift the embargo on interpersonal comparisons of utility for a moment, she gets more from a concert with him than he from a bullfight with her; and the *total* utility of the former is more than that of the latter. Now is the concert salient? The answer is still No. If these can be motivating factors, then they are already included. The theory has no view on what motivates us or what should motivate us. It insists merely that Jack's motives are Jack's and Jill's are Jill's, with each being able to be represented as pay-offs to that player.

Philosophical doubts begin to stir. But I am sure of my ground, not least because of an arresting bit of argument in an elegant recent survey article on game theory by Binmore and Dasgupta,[2] whose details I shall adapt to Jack and Jill. It is some months later and their relationship has developed. He has become more involved with her, she more nuanced about his place in her tastes and feelings. By now she would be equally delighted to go to a concert with him or (surprisingly, but let it pass) a bullfight without him. She would be less delighted but equally content to go to a bullfight with him or a concert without him. He, more simply, always prefers her company, so much so that he now most fancies a concert at her side. Here they are in *game 3*:

---

[2] K. Binmore and P. Dasgupta, 'Game Theory: A Survey', in their edited collection, *Economic Organisations as Games* (Oxford: Basil Blackwell, 1986), pp. 1–45, and especially p. 17 for a discussion of the point about 'trembling hands'.

		Jack	
		Bullfight	Concert
Jill	Bullfight	2,2	5,0
	Concert	2,1	5,5

game 3

Now surely the concert is salient? Certainly one might think it the rational choice for Jack, with the bullfight no longer worth more than a possible '2' and noticing that she is musically inclined. But, even if that case could be made, it still would not be Jill's rational choice. She is austerely indifferent. If he is going to choose the bullfight, she will score '2' whatever she does. If he is going to choose the concert, she will score '5' whatever she does. She has no reason to prefer the concert, at any rate in the absence of communication.

To prove the point let us introduce an element of what game theorists engagingly call a 'trembling hand'. This is the faint possibility that even a rational player might play a non-equilibrium strategy by mistake, for instance by pressing a wrong button in a fit of hiccoughs and thus starting a nuclear war. Consider *game 4*, where Jack just might step under a passing car on the way out of the hotel and be crippled for life — a third course of action, so to speak, which he would take only by accident, through male clumsiness.

		Jack		
		Bullfight	Concert	X
Jill	Bullfight	2,2	5,0	−99.9, −1000
	Concert	2,1	5,5	−100, −1000

game 4

If, in that case, Jill would rather have gone to the bullfight (in pious memory of old days perhaps), then this is what she will do, however small the risk of the accident. Being otherwise indifferent between her two options, she will be swayed by the slightest disturbance of the balance. The bullfight becomes her dominant choice, thus prompting Jack to aim for that equilibrium.

I call this an arresting bit of argument, because it took me some effort to swallow it. Offhand, it sounds ridiculous. It is bad enough to be told that concert-concert (5,5) is not salient in *game 3* with hyper-rational players, and worse to have (2,2) become salient in *game 4*. But the argument is formally correct. All motivating elements are already included in the pay-offs. Jill's prospects improve by a tiny fraction, if she chooses the bullfight in *game 4*. For a hyper-rational player this breaks the equipoise.

To swallow the argument, however, is only to shift attention to its assumptions. I shall argue next that it requires an ideal-type world crucially unlike our own. There are three groups of assumptions to think about, all of which can conveniently be extracted by noting how the pay-off numbers function.

Firstly, the pay-offs express an assumption that all outcomes are grist to the mill of an agent's preferences. This is strong and silent stuff. It is strong in that Jack can compare not only bullfights with concerts but also guns and butter, death and dishonour, the scratching of his little finger and the destruction of the whole world. It is silent in that 'utility' is a purely notional comparator. Originally it was supposed to measure 'a psychic flow of inner satisfaction' (in Irving Fisher's phrase). But nowadays Jack needs merely to know somehow what his ranking of any pair of outcomes is. Although I am using the utility numbers merely as a graphic device, even ordinal comparisons involve complete and consistent self-knowledge of, I am bound to say, a mysterious kind. Let us mark the silence about the measure and what it refers to by noting a lack of *grit*, meaning something for the preference order to bite on. To dispense with all grit is to assume that there is no problem in assimilating motivating elements which can include Jack's aesthetic response to the music on the concert programme, his pleasure in Jill's pleasure, his recognition of her as a person whose preferences are to be respected and his moral reflection that concerts avoid some ethical questions set by the cruelties of the bullfight.

Secondly hyper-rational agents have complete information about themselves and others. There is a complete lack of *dust* in the hyper-rational air. In the limiting case Jack and Jill are so wired into each other's inner states that, were they computers, we would be asking whether they were two computers or one with two interconnected routines. Thus, when I set my chess programme to the 'Self-play' mode, I can pretend that black is playing white, as if one was Jack and the other Jill; but the pretence is possible only because, when Jack really plays Jill, neither is completely transparent. Hyper-rational agents can set no traps (since they would be instantly detected) and can make no mental mistakes. Since only a random intrusion can prevent the result implicit in the logic of the game, victories are empty and the game pointless. Similarly, if all bargains are not sealed instantly without incurring any costs and if conflicts ever occur despite deadweight losses, one asks why. One's left hand does not dispute with one's right hand over which shall reach for a cooling drink which is meanwhile evaporating.

A 'trembling hand' throws a pinch of dust into the air, thus letting us see how hyper-rational agents deal with a whiff of uncertainty. But it introduces a physical disturbance which distracts both players equally. If we want to insist that the players need to be separate persons, then I think that they must be mentally opaque to each other. That finally requires a moral psychology in which agents are not, so to speak, patients, as I shall argue presently. For the moment, however, I introduce the point schematically. It has been assumed so far that Jack and Jill hold the same theory of motivation, so that each can be sure how the other will act, given the preferences of each and the information available to both. What theory is that? Presumably it is the one implicit in the basic definition of rational choice. But what if this theory of motivation can be challenged? Then there is room for doubt about what other players will do. As soon as this doubt is possible, however, it infects the certainties which have been giving us pause. For instance in *game 3* Jack was sure that Jill was indifferent between bullfight and concert, and sure in *game 4* that she would choose the bullfight. If he cannot count on her deliberations proceeding as the theory prescribes, then his rational choice becomes indeterminate.

To put flesh on this schematic line of thought, notice, thirdly, that there is no *glue* in the hyper-rational world. I mean that the concept of rationality is thoroughly instrumental, requiring that a rational agent is always ready to switch strategy for the sake of a marginal gain in his overall utility. This sets an intriguing problem of what Derek Parfit aptly terms 'bias to the near',[3] which is troublesome even for a solitary agent planning his future and poses a radical threat to Jack's relations with Jill. Notoriously Jack will sometimes do better for himself, and he and Jill will both fare better, if he can overcome the built-in bias to whatever best satisfies *his* current preferences. But, although this suggests suspending immediate considerations, it does so only for tactical reasons and only for as long as those reasons hold. Any commitments therefore seem to bind him only *pro tempore*. To make them more durable, he needs to hold a *moral* theory as part of the theory of motivation which he has in his head and Jill needs to know what it is. Thus game theory seems to involve a moral psychology which goes beyond the standardly recognized assumptions about motivation.

Here lies water so deep that game theorists may hesitate. Their best retort, I think, is that the point stands only if we demand that individually rational choices shall not sum to Pareto-inferior (or

---

[3] In his *Reasons and Persons* (Oxford: Oxford University Press, 1984).

'collectivity irrational') outcomes. Why assume that rational agents do not cut off their noses, thus spiting their collective and several faces? Yet, it is at least tempting to accept the point. Then the question becomes what moral theory to ascribe to rational agents. The current line is usually that a contractarian ethics fits the bill. But whether it truly will is deep water indeed.

The issue comes to a head in *game 5*, where Jill's choice of the bullfight is dominant and Jack, knowing this, chooses the bullfight too:

		Jack	
		Bullfight	Concert
Jill	Bullfight	−2, −2	6, −3
	Concert	−3, −3	5, 5

game 5

(The minus signs are there to make the point graphic, even if they raise a query which I shall not linger on). They spend a rotten evening, reflecting how much both would rather that both had opted for the concert. They surely need something to distance themselves radically from the utility numbers. But, so far at least, a hyper-rational world without grit, dust or glue has them in thrall.

To save time I shall jump straight to the eventual question. What kind of grit, dust or glue would get them to the concert? An obvious diagnosis is that Jack and Jill are too instrumental for their own good. But this way of putting it suggests that it would pay them, instrumentally speaking, to be less instrumental. If so, the snags are prone to recur at the second order. So I shall cast around further afield by trying out two more radical challenges to the Humean philosophy of mind assumed so far, one broadly Wittgensteinian, the other broadly Kantian. By 'Humean' here I mean partly the role given to desire (or preferences) as the motor of action, with belief (or information) relegated to being the slave of the passions, on the ground that 'reason alone can never be a motive to any action of the will' (*Treatise* II.iii.3). This will be the target of a Kantian riposte presently. But I also mean the wider picture of action as an effect whose cause is a mental state composed of desires and beliefs. Wittgensteinians will have none of that.

When game theory abstracts from social life to a 2 × 2 matrix, it skins off the social context so as to isolate a universal core. All social embodiments of the same matrix are game-theoretically equivalent. It does not matter whether the agents are persons or, for instance, firms or nations, nor whether the options are ways of spending an evening or ways of conquering the world. This gives the approach its power but also leaves it vulnerable. Supposedly Jack and Jill have been a mere illustration for the benefit of those who are not quite at home with pure abstractions. In fact, I submit, an implicit and specific social context has been doing a great deal of work. Jack and Jill already have some kind of relationship relevant even to their one-shot games and exemplifying not the 2×2 matrix but a common pattern of such relationships located in what might be called a culture or form of life. The matrix is a snapshot of a situation reached by a trajectory and fraught with expectations. If this matters, and especially if it matters how the situation came about, then different exemplifications of the game need not be equivalent after all.

The players sound like Brits. The place sounds like Spain. Even if they are strangers to each other, it may matter who they are and why they are there. Normative expectations surround the meeting of strangers, depending on whether they are tourists or immigrants and how each is socially located at home as well as abroad. Each dresses, moves and speaks as if flying a string of flags for the other to decode. These signals may be conscious or unconscious, easy or hard to read, but they are certainly visible to an informed eye. Each discloses something about class, income, occupation, character and attitudes, thus staking out the normative ground for their cultural or sexual encounter. Their options have been listed as 'bullfight' and 'concert' but these are very curt specifications of options saturated with many species of meaning. Game theory seems at first sight to declare that all this is irrelevant, since the pattern of the pay-offs and common knowledge that the players are rational is enough to go on. But, alternatively, perhaps it includes some typified reckoning of the context and character of the players. In that case the matrix is not silent where it seemed silent at first.

Look again at the components of the pay-off numbers. Each number is a homogenized sum of assorted satisfactions, for instance Jack's tastes, his pleasure or pain in Jill's responses, which may include his pleasure in her pleasure in his pleasures, his sense of what is proper or right, and so on. Round each component hover subjunctive conditionals, to do with what difference it would make if something in the situation were different. For example, if Jill became more disgusted by bullfighting, that would lower Jack's utility for bullfights, unless he became less enamoured of Jill. Such conditionals do not operate in a human or cultural vacuum. Their context is highly relevant as soon as either player needs to take them into account. Suppose, for the sake of argument, that the 3 utils, which Jack gets from their both going to the bullfight in the original *game 1* includes -1 because Jill is queasy about its cruelty (offset by her finding it exciting, yet incompletely, in that her '1' in the bottom left box has something to do with her satisfaction that he is getting a dose of higher culture). There is a hard question about how he knows that '−1' is a reliable figure and it leads us to ask how *she* knows. The whole matrix is shot through with huge assumptions about the power of introspection. As soon as introspection is challenged, the conditionals cease being merely secondary. The mental states, which were supposed to sit still to be photographed for the matrix, are replete with guesses about likely responses to shifts in the world where players interact.

This line leads to the idea that Jack and Jill are playing 'games' of another character altogether. There will no longer be some single point which all games have in common, like the maximizing of an individual's pay-off. Instead there will be only criss-crossing family resemblances, sharing only the loose family tie that all are constituted and regulated by rules governing the way to go on. Interplay becomes, in some sense, prior to play, rather as tennis is not the meeting of persons, who have separately chanced upon the same individual and solitary pastime, but a communal activity which makes sense of individual shots. This is Wittgenstein territory and, by saying so, I hope to be excused an exposition. Instead I shall pick out two large implications and raise one large query.

The first implication is that we cannot isolate a pay-off matrix common to games with different contexts. The mating game which Jack and Jill play and the diplomatic game which arms negotiators play cannot both be fully captured by, say, the Prisoner's Dilemma. For, even though one might plausibly wish to say of each set of players that they are stuck with self-defeating dominant strategies, it matters now how they came to be stuck. This is what I had in mind when speaking of a snapshot of a situation reached by a trajectory and fraught with expectations. 'Trajectory' is a metaphor for the forward integration of context and moves. The feasible moves are permissible options for how to go on and their significance is a function of the context conceived as rules of that particular game. That is one reason why introspection has become suspect. There is no reason to credit a chess player with an adequate knowledge of what he is doing, just because he makes introspective claims to himself. We must first have reason to be sure that he understands the public game properly. If game theorists reply that even in one-shot games players can be credited with external information, the retort is that this kind of information is, crucially, intrinsic to the description under which players consider their situation.

Secondly 'expectations' in games from a Wittgensteinian family are not (or not only) predictions, as in game theory. *Normative* expectations are involved. For instance Jack and Jill will have *entitled* each other to expect typical responses and their interaction is charged with a moral vocabulary. Even in what looks like a genuinely one-shot game, where they have met as strangers intersecting only for a single evening, they bring normative conventions to the brief encounter. Otherwise they will not succeed in interacting, even if they happen to occupy the same place at the same time. The normative expectations surrounding a one-night stand are not those of a disarmament conference. Hence the figures in *game 3* may not signify indifference on the part of the row-player in all contexts. The convention govering how the sexes take their leisure may prescribe that Jill resolves ties in Jack's favour.

The standard game-theorist will, of course, still dispute all this. If Jill turns out not to be 'indifferent', then the utilities in the pay-offs have been wrongly ascribed. But this is not compelling, if the complaint is that the pay-off numbers are overloaded because they are an attempt to represent as a single net sum what remain complex hypotheticals with normative components. Meanwhile the contextualizing of games does at least suggest fresh ideas about the relation of game-theory to morality. One favourite way of getting Jack and Jill into the right box together and out of a Nash-equilibrium which is inferior for both is, at least in games belonging to a series, to have them make one another promises. A familiar snag is that promises made for instrumental reasons should rationally, it seems, be broken whenever it pays. Perhaps this snag does not bite where a game is the extrusion of a normative context governing both moves and motives.

That thought, however, also introduces my large query. In one reading of Wittgenstein, the inner mental states of Humean moral psychology are so radically externalized that moves and motives can be identified solely through the context, as if to know the rules fully was thereby to know how to go on. I shall take it, however, that the players are to be left some latitude and the rules are to be construed in a more constructivist way. The model game is not maypole dancing, where the aim of each is exhausted by the collective aim of making a pretty pattern, but tennis or chess, where each is trying to win. But then we have a large query about the relation of players to rules. Specifically, are we to see the kind of moral commitment, which might serve Jack and Jill well, as a demand by the context or as an interpretation by the players? If the latter answer seems to reinstate the snag by restoring Jack and Jill their individuality, the former seems to swallow them up.

Anyone stuck by the force of this dilemma might care for a Kantian line. Some features of the context are re-internalized but as demands of a Reason which is no longer the slave of the passions. Ideally rational agents, playing a game where moral choices must be made, will choose from 'the moral point of view' — the universal, impartial, impersonal standpoint which Kant advocated. What started as instrumental hyper-rationality now emerges as moral autonomy. Even the fact that Jack and Jill both do better for themselves becomes an incidental, if gratifying, by-product of the fact that each has chosen what is right. The relation of players to rules is that autonomous players follow the guidance of external reasons in their internal deliberations.

This line requires an explicitly two-tier moral psychology, whose lower tier is broadly Humean and whose upper tier is firmly not. The lower tier is one of hypothetical imperatives, where a rational agent can proceed by calculating the most effective means to satisfy desires. Lower tier precommitments can be rational, as when Ulysses has himself bound to the mast so as to hear the sirens and survive. But they are instrumental devices and psychologically artificial. If the cord binding Ulysees is internal, in that it consists only of a promise given to himself or others, then he can always untie the knot, and will untie it as soon as inclination or prudence demands. By contrast the imperatives of the upper tier are categorical. They forbid treating other persons as means to one's ends and they guide an autonomous agent independently of his sentiments or calculations. Motivation on the upper tier is netted into a community of persons, a kingdom of ends or *Rechtsstaat*, so that moral precommitment becomes binding for as long as autonomy lasts. Rules become full-bloodedly normative, providing that they satisfy these moral conditions, because they constitute the form of life which a moral agent embodies.

I shall not try to justify a Kantian ethics here. But I do want to suggest that game-theorists, whose moral psychology cannot cope with a Wittgensteinian revolution, might find succour in Kant. The crux concerns the possibility of drawing up pay-off tables which are neutral between socially different contexts. The Wittgensteinian objection is that there is no neutral currency of satisfactions for assessing actions which belong to different practices. Underlying it is a suggestion that to be rational is to follow an appropriate rule appropriately, rather than weighing options as if in the omnicompetent pans of a pair of scales. To accept it would be catastrophic for game theory, as it destroys the whole basis of utility theory. At the same time, however, it threatens to inject so much grit, dust and glue that agents lose all distance from their situation and become cultural dopes. That can hardly be the right reading of Wittgenstein but, whatever is, game theory needs to hang on to agents who have a rational will operating at a distance from context. Can Wittgensteinians somehow oblige?

Meanwhile, Kantian moral theory is less catastrophic. It assigns an agent's 'ethical preferences' to an upper tier, provided that they are genuinely moral as certified by their passing the universalizability test, and bids him calculate from a moral standpoint. This certainly distances him from his situation. It thereby offers game theory a two-tier motivational structure, in which preferences over preferences operate differently from preferences over outcomes.

This seems to me a promising advance. At the same time, however, there are snags. Any inflexible maxim like 'Do what others would most prefer' is liable to recreate the original problems with, so to speak, the colours reversed. There are Prisoner's Dilemmas for altruists as well as for egoists. Also a maxim commanding unselfishness can paralyse everyone, if everyone follows it. (Try cooking dinner for a group of relentless altruists with different tastes.) More seriously still, even if some choices do need to be placed morally off limits, others surely involve a moral or upper-tier judgement which is sensitive to consequences, including what other players are likely to do upon reflection. Otherwise ethics loses touch with the domain of rational interaction. Game theory has nothing to learn from the moral will of a rhinoceros. Thus there is no simple Kantian magic on offer. We certainly need a glue resistant to bias to the near; but one less rigid than Kantian superglue.

A test question is what the rational choices are in *game 5*, where Jill's apparently dominant choice steered Jack, and hence both of them, to an outcome inferior for both. On the one hand the 'collectively rational' outcome is the concert (5,5) and it is tempting to make it accessible. On the other hand received opinion about the one-shot Prisoner's Dilemma is that defection is the rational choice, and I am convinced by it. So I shall end by suggesting that rational choice can depend on the context as well as the utilities. Revert to the nursery rhyme and the game which faces Jack and Jill, when someone confronts them with the broken pieces of the china pail, which should never have been taken up the hill, and asks who is to blame for this naughtiness. If both own up, each will lose a month's pocket money but honour and justice will have been so well served that both emerge with credit. If neither owns up, they will not be believed, will still have to pay and will emerge in disgrace. If Jill alone owns up, she will forfeit two month's pocket money but without dishonour, whereas Jack will feel simply terrible. If Jack alone owns up, however, Jill, being nastier and naughtier than he, will be thoroughly pleased with herself. Quantifying notionally, we get *game 6*:

		Jack	
		Dishonesty	Honesty
Jill	Dishonesty	−2, −2	6, −3
	Honesty	−3, −3	5,5

game 6

Although *game 6* has the same numbering as *game 5*, the labels are significantly different. Thus the crucial '6' in the top right box records Jill's 'preference' for a ride on Jack's honesty. But it suppresses some upper-tier questions about trajectory and expectations. How is that outcome to be described? It is one which is, for instance, unfair, exploitative and lands Jack with the misplaced opprobrium of leading his sister astray. Is this the description under which '6' is the utility figure correctly recorded for Jill? If not, is it, none the less, the description which should be used by a rational agent deciding whether she has better reason to be honest or dishonest? When Jill reflects that 'if Jack fails to own up, I am rational not to own up' and reflects that 'if Jack owns up, I am rational not to own up', are these reflections truly parallel?

Such questions prompt me to challenge the assumptions which connect individual pay-offs to reasons for action. A direct connection requires agents who are, in the ideal-type case, transparent maximizing machines. But there are other ways to conceive of persons and, correspondingly, other ways to read the table of utilities. I have gestured both to a Kantian conception of persons as ends-in-themselves, with whom a truly rational agent will deal by seeking the impersonally right outcome. I have also gestured to a more situated concept of persons as bearers of locally established norms and practices, who will do whatever best accords with the prevailing rules. A neat way of contrasting these alternatives might be to invoke Bernard Williams's distinction between 'thin' concepts like 'good', 'right' or 'fair' and 'thick' concepts like 'honourable', 'kind' or 'considerate'. That, of course, merely sets more problems, when we ask whether concepts like 'honest' or 'reasonable' are thick or thin. Yet this seems to me the right sort of question to ask about the 'persons' whose games are to be illuminated by a theory of games.

Amid these deepening complexities game-theorists may lose patience with philosophers and withdraw into an ideal-type world where utilities are unambiguous reports of simple preferences and motivation stems directly from them after discounting for risk. But, if so, I would be sorry. Game theory relies on a double abstraction from our world of laughter and forgetting. It first idealizes persons as rational economic individuals and then idealizes their interaction for a limiting case, where there is no grit, dust or glue. The second idealization is clearly prescriptive and, supposedly, can be made usefully so by injecting measured amounts of grit, dust and glue and then reworking the prescriptions. But, in bringing the ideal-type nearer to earth, the theory unsettles the first step, because it needs agents distanced from their preferences. That calls for a philosophy of mind where expected utilities do not form a logical bond between moves and motives. The heart has its reasons, but not the kind which expected utility theory knows all about. The head has its reasons too, but not always the kind that can be priced in utilities. We are looking for a philosophical psychology where persons have grit in their motives, dust in the spaces between them and glue in their relationships.

Yet Jack and Jill are still playing games. The metaphor remains illuminating. There are pails of water to fetch, which one cannot carry alone and two cannot manage without a concern for the common good. If game theory cannot see them safely home with its standard assumptions, I hope that philosophy can help find some better ones.[4]

School of Economics and Social Studies,
University of East Anglia, Norwich NR4 7TJ

[4] I would like to thank Robin Cubitt, Shaun Hargreaves-Heap, Bruce Lyons and Robert Sugden warmly for their helpful comments on an earlier draft.

FILLING IN SPACE

By SIMON BLACKBURN

WHY do people think that dispositions must have categorical grounds underlying them? Well, the clock tells the time because there is such-and-such an arrangement of little bits inside it; Sandy barks because her vocal chords vibrate; the light glows because electrons whizz around in its filament. The explanations are excellent, but do they illustrate the doctrine?

They do if they bring us to categorical grounds that indeed underlie the dispositions. But they do not if they bring us to something else. For instance, they might only bring us to a point where the possession of some disposition by something is explained by the possession of other dispositions by the same or different things. Or they might only bring us to the instancing of a power (disposition) at some region of space explained by the instancing of some other power at some related region of space.[1]

When we think of categorical grounds, we are apt to think of a spatial configuration of things — hard, massy, shaped things resisting penetration and displacement by others of their kind. But the categorical credentials of any item in this list are poor. Resistance is *par excellence* dispositional; extension is only of use, as Leibniz insisted, if there is some other property whose instancing defines the boundaries; hardness goes with resistance, and mass is knowable only by its dynamical effects. Turn up the magnification and we find things like an electrical charge at a point, or rather varying over a region, but the magnitude of a field at a region is known only through its effect on other things in spatial relations to that region. A region with charge is very different from a region without: perhaps different enough to explain all we could ever know about nature. It differs precisely in its dispositions or powers. But science finds only dispositional properties, all the way down.

---

[1] To avoid clutter I should say that from now on I shall talk indifferently of powers, dispositions, and 'counterfactuals' to describe the features on the non-categorical side of the fence; it is their contrast with the other side that matters, not possible differences between them.

Evans talked of a prejudice against bare ungrounded dispositionality, 'equally offended by the idea of two places alike in what occupies them between visits, yet of which one is such that if one goes to it, one will have certain experiences, and the other is not'.[2] He wanted as well a 'relatively abiding property' that fills out the intervals in the exercise of dispositions. Physics can only give him this in a backhanded way that will not satisfy the demand. An electrical field can abide, certainly, but that just means that there is a period of time over which various counterfactuals are true. It does not give us reason to think of a different property, quietly persisting, as it were, even when the disposition is not exercised.[3]

Perhaps I have been unduly verificationist. True, you might say, things like fields and masses are known by their effects on other things. But this should not prevent us thinking of them as *in themselves* categorical. There will be a categorical ground, G, for the (multi-track) disposition D whereby we know of mass or charge. It will be in virtue of the instancing of such a G that an object has the mass that it does, or a region of space the charge. When we think this, I believe we surreptitiously imagine an improvement in science that would enable us to identify G: a new theory about something true of charged regions of space, for example, rather like the molecular theory of gases. But this road only leads to the same place. Just as the molecular theory gives us only things with dispositions, so any conceivable improvement in science will give us only a better pattern of dispositions and powers. That's the way physics works.

Is it the way it has to work? I believe so. A quick route to this conclusion is to see the theoretical terms of a science as defined functionally, in terms of their place in a network of laws.[4] A slower route is to reflect on what is needed from physical thinking. What is needed is the use of concepts — energy, temperature, entropy — that cover changes of state, permitting the formulation of conservation laws. Such concepts in effect tell us what is the same about a changing system, in terms precisely of its powers and dispositions.[5]

---

[2] Gareth Evans, 'Things Without the Mind', in *Philosophical Subjects*, ed. Zak van Straaten (Oxford: Oxford University Press, 1980), p. 102.

[3] Cf. Strawson's reply to Evans, p. 280. 'If it seems true of the sensory properties in general that they all dissolve together, under reflective pressure, into dispositions, this seems even more certainly true of the "physical" properties which are held to constitute their categorical base.'

[4] David Lewis, 'How to Define Theoretical Terms', in his *Philosophical Papers*, Vol. 1 (Oxford: Oxford University Press, 1983).

G will remain, therefore, entirely beyond our ken, a something-we-know-not-what identified only by the powers and dispositions it supports. And then the possibilities multiply. Perhaps there is a possible world just like ours, not only in surface appearance, but in all that physics could ever discover, in which the dispositions have a different categorical ground, G′. Perhaps in our own world G′ supports dispositions on Mondays and Wednesdays, while G supports them on the other days. But the real problem arises with the nature of the underlying and support that G provides for D. Is it logically necessary that G supports D? Presumably not — what could logic have to say about it? And since we know nothing of G it can hardly be a priori in any other way that it does what it does. Presumably then there is a law whereby G supports D and this law imputes a power to G. So it ought to need a separate categorical ground, G*, it being in virtue of G* that G gives rise to D in the worlds that obey this law. But then the power of G*, to bring it about that G gives rise to D, will itself need a ground, and so forever. To stop the regress we need a brute or bare power without a categorical ground: better in that case not to insist on grounds in the first place.

Are we in danger of proving too much? It seems as though we have excised categorical properties from nature altogether, leaving only features that, as Russell said, are each other's washing.[6] The problem is very clear if we use a possible worlds analysis of counterfactuals. To conceive of *all* the truths about a world as dispositional, is to suppose that a world is entirely described by what is true at *neighbouring* worlds. And since our argument was a priori, these truths in turn vanish into truths about yet other neighbouring worlds, and the result is that there is no truth anywhere.[7] Here is a good analogy. Early philosophers such as Whitehead and Collingwood, reflecting on modern physics saw that it resolved substance into function, and drew the consequence that there is no such thing as a state of nature at a literal timeslice: processes take time. 'There is no nature at an instant'.[8] The present problem is that processes take possibilities as well as actuality, so there is no nature at an actual world.

---

[5] I tell this story at greater length in 'Losing your Mind: Physics, Identity, and Folk Burglar Prevention', forthcoming.

[6] *The Analysis of Matter* (London; Kegan Paul, 1927), p. 325. 'There are many possible ways of turning some things hitherto regarded as "real" into mere laws concerning the other things. Obviously there must be a limit to this process, or else all the things in the world will merely be each other's washing'.

[7] I have subsequently found this point made in very similar terms by Howard Robinson, *Matter and Sense* (Cambridge: Cambridge University Press, 1982), Chapter 9. Robinson uses it to attack any conception of matter, whereas I leave the issue in Hume's hands rather than Berkeley's.

In Lewis's exploration of these issues, categoricity comes in the 'Humean mosaic' or pattern of 'perfectly natural intrinsic properties which need nothing bigger than a point at which to be instantiated'.[9] It seems as though we need them, but it now also seems as though we cannot have them — out best physical understanding of the world gives us no conception of what they might be. But the mention of Hume is suggestive. We can think of the Humean mosaic in experimental terms: a colour here, a tactile sensation there, a sound somewhere else. Categoricity in fact comes with the subjective view: there is nothing dispositional, to the subject, in the onset of a pain or a flash in the visual field. Such events come displayed to us as bare, monadic, changes in particular elements of experience. In this perspective a change in perceived colour is as categorical as a change in shape or a twinge of toothache, even if from the objective standpoint, 'all that goes on' when such changes occur is that a change of functional (dispositional) state arises, the subject being disposed to act and think differently as a consequence of changes in the dispositions of surrounding things.

The trouble now is that such events, conceived of as categorical, play no role in a scientific understanding of the world; they certainly do not serve to ground anything. The question remains whether we can live with a concept of objectivity that leaves us only powers without end, and escape from the discomfort this causes by somehow retreating to a subjective point of view. Strawson counsels that we can, self-consciously shifting perspectives to avoid the contradiction of thinking of the same event as both categorical and dispositional.[10] But the problem remains that this gives us no help in understanding what, except counterfactuals, is true of the objective order of nature, unless, heroically, we see that order as a kind of construct from the categorical point-instances of properties available to the subjective view — a kind of neutral monism. It almost seems that carelessness and inattention alone afford a remedy — the remedy of course of allowing ourselves to have any idea at all of what could fill in space.

Pembroke College,
Oxford OX1 1DW

---

[8] See for example, R. G. Collingwood, *The Idea of Nature* (Oxford: Oxford University Press, 1945), Introduction, Sections 4 and 5. The saying is quoted from Whitehead, *Nature and Life* (Cambridge: Cambridge University Press, 1934), p. 48.

[9] David Lewis, *Philosophical Papers*, Vol. II (Oxford: Oxford University Press, 1986), Introduction, p. x.

[10] Op. cit., p. 280-1.

WHY SUPERVENIENCE?

By DAVID PAPINEAU

I

IT seems plausible that such categories as the psychological, the biological, the meteorological, the chemical, and so on, *supervene* on the physical, in the sense that two situations cannot differ psychologically, or biologically, or whatever, without differing physically.

In this paper I want to show *why* this doctrine is true. By doing so I hope both to clarify the doctrine and show how it can be defended against a common objection.

For brevity I shall concentrate on the relation between the psychological and the physical. But the analysis will be of general significance, as the structure of the overall argument will make clear.

Consider first the following argument. Mental occurrences have physical effects. For example, Eric Bristow's desire to score thirty-two at darts causes him to hit double sixteen. But such physical effects are also attributable to physical causes. The trajectory of Eric Bristow's dart is also caused by the arrangement of neurones in his brain and his consequent bodily movements. So unless we want to say that such physical effects are overdetermined by two separate causes, which we clearly don't, we need somehow to view the mental cause and the physical cause as the same cause. (Cf. Peacocke [1], Ch. III.3.)

There are various possible ways of instituting the required relationship between mental and physical causes. If we think of events as bare particulars, we can say that each particular mental cause is the *same event* as the relevant physical cause (cf. Davidson [2]), whereas if we think of events as instantiations of properties (cf. Kim [3]), we will prefer to say that that the particular mental cause is 'realized by', or perhaps 'constituted by', the physical cause. Alternatively, we might hold that causation is not a relation between events at all, but rather a relation between facts (cf. Mellor [4]), and so take the above argument to show instead that particular mental *facts* are identical with, or realized by, or constituted by, physical *facts.*

Whichever of these options we adopt, we shall have a sense in which mental particulars have physical features as well as mental features.[1] But such a doctrine of 'token identity' is not yet supervenience. Supervenience requires not just that mental particulars have both physical and mental features, but more specifically the asymmetrical doctrine that those physical features *determine* those mental features.

---

[1] We need not suppose that these physical features are exhausted by what is in the head. However, the need to go extra-cranial involves questions about the causal significance of intentional contents, and I shall not pursue this difficult topic here.

Let us look more closely at the assumption on which the above argument for token identity is based, namely, the assumption that all physical effects, including those with mental causes, also have physical causes. I take it that the reason this assumption recommends itself to us so strongly is that we are committed to the *completeness of physics*: all physical events are determined (or have their chances determined[2]) entirely by prior physical events according to physical laws. Indeed some philosophers have argued explicitly from the completeness of physics to versions of token identity. (Cf. Lewis [5]; Smith and Jones [6], pp. 57–9.) What I want to show in this paper, however, is that the completeness of physics yields an argument, not just for token identity, but also for the asymmetrical thesis of the supervenience of the mental on the physical.

In order to do this it will be convenient first to introduce an intermediate thesis, which I shall call the principle of *no independent causal powers*: whenever any mental event causes another event, it does so entirely in virtue of its physical features, in the sense that in the context the effect is fixed entirely by the physical features of the mental cause.

It is easy enough to see that supervenience will follow once this principle of no independent causal powers has been established. For this principle tells us that two situations which differ mentally, but without differing physically, will have exactly the same effects, in virtue of their physical similarity. But a difference which cannot possibly show up in any differential effects is surely no difference at all.[3] So in order for two situations to differ mentally, they must differ physically as well.

II

The task, then, is to show that the completeness of physics implies the principle of no independent causal powers. Consider first the case where a mental cause has a *physical* effect, as for instance when Eric Bristow's desire causes his dart to hit the double sixteen. The completeness of the physical tells us that this physical effect is entirely determined by prior physical events according to physical laws. So it follows immediately that, when Bristow's desire plays a part in determining the dart's trajectory, it must do so in virtue of its physical features.

---

[2] In what follows I shall simplify by ignoring indeterminism. The argument can easily enough be generalized to cover indeterministic effects.

[3] Colin McGinn ([7], p. 29) argues that, if we deny the supervenience of the mental on the brain, then we will be committed, absurdly, to behavioural differences without physical causes. In effect I am generalizing this argument, so as to block the thought that the mental might manifest its independence from the physical by differential *non*-behavioural effects.

What about the case when a mental cause has a *mental* effect, as when Eric Bristow's *mental arithmetic* causes his *desire to score thirty-two*? But such mental effects will generally themselves have physical effects in turn, such as the dart actually hitting the double sixteen. In which case the original cause will cause a physical effect *via* causing the intermediate mental cause, and we will be able to apply the argument of the previous paragraph in stages. By the completeness of the physical, the role of the intermediate cause (the desire to score thirty-two) in fixing the eventual physical result (the dart hitting the double sixteen) must depend on the physical features of that intermediate cause. But then those physical features of the intermediate cause must depend, by the same argument, on the physical features of the original cause. And then, if we want to say that the original cause produces the eventual result *via* the intermediate cause, as surely we do, we will need to attribute the intermediate cause as such, and not just its physical features, to the physical features of the original cause.

It might seem as if this argument for the thesis of no independent causal powers begs the question, by focusing exclusively on *physical* laws. Suppose there are also laws according to which certain *mental* antecedents suffice for certain effects. Then doesn't it immediately follow that those mental antecedents have those effects in virtue of their *mental* features, as much as in virtue of any physical features from which those effects may also follow by physical law?

I am happy to accept that such mental laws may exist.[4] And, of course, if they do, then the effects they cover will be fixed by mental antecedents as much as by physical ones. The crucial point, however, is that, even if some such mental laws exist, they won't pervade the world in the way that physical laws do.

Note that the mental, unlike the physical, is *not* complete. It is not the case that all mental events are determined, or have their chances determined, entirely by prior mental events according to psychological laws. After all, there are some mental events which don't have any mental antecedents at all, as when you sit on a drawing pin and feel a pain.

And so there is no question of arguing that *all* effects of physical causes are fixed by the mental features of those physical causes, in the way that we argued that all effects of mental causes are fixed by the physical features of those mental causes. Even if we accept the not entirely implausible premiss that every physical cause eventually yields some mental effect, we won't be able to conclude that it does so in virtue of its mental features, since it mightn't have any mental features to start with, as when the drawing pin causes the pain.

---

[4] Question: if there are such laws, and supervenience holds, must those laws be reducible to physics? Answer: not necessarily, but such an absence of reducibility requires the presence of natural selection. (See Papineau [8].)

So the overall situation is as follows. Whenever a mental cause has any effect, that effect is fixed in the context by the physical features of that mental cause. By contrast, only some effects of some physical causes are fixed by mental features of those physical causes. Which is why mental differences without physical differences cannot show themselves, whereas physical differences without mental differences can.

III

There is both a strong and a weak version of supervenience. The weak version says that, within any possible world, two situations cannot differ mentally without differing physically. The strong version says that two physically identical situations will be mentally identical even if they are in different possible worlds. (See Kim [9].)

The argument of the last section supports a version of the strong version of supervenience: physical characteristics fix mental characteristics at least across all *naturally* possible worlds in which mind and matter interact. This is because the argument for supervenience was simply that differential effects, of a kind which might manifest a mental difference among physical identicals, are impossible as long as mental occurrences have physical effects and the law of physics hold good.

It might strike some readers as odd that I want to offer such a complex argument for something as intuitively obvious as supervenience. However, by way of corroboration of my analysis, it will be illuminating to consider what would be intuitively obvious to somebody who had a *phenomenalist* metaphysics, rather than the metaphysics I have presupposed so far.

Phenomenalism is committed to the completeness of the *phenomenal*, rather than the completeness of the physical: a phenomenalist will hold that all phenomenal facts are entirely determined, in so far as they are determined at all, by prior phenomenal facts, according to laws governing the occurrence of phenomena. Because of this it will seem obvious to a phenomenalist that *physical* events have no causal powers apart from those that derive from their *phenomenal* constitution. And because of this the phenomenalist will conclude that the *physical* supervenes on the *phenomenal*, rather than conversely (for how could a physical difference without a phenomenal difference ever show itself?)

Let me conclude by addressing a common objection to the supervenience of the mental on the physical. This objection poses a dilemma for the notion of *physics* involved.

If the past form of physical theories is anything to go on, then currently accepted physics is false. So the mental can scarcely supervene on currently accepted physics.

On the other hand, if there is a true future physical theory, we don't yet know what concepts it will use. So what grounds can we have for holding that the mental supervenes on the true physics?

The analysis of this paper offers a way round this dilemma. We can use the notion of *completeness* to give substance to the idea of the true physical theory. The argument of this paper hasn't just appealed to intuitions about physical concepts, but rather argued from the completeness of physics. So provided the true physics is complete in the relevant sense, the mental will supervene on it. In effect, we could say that, for the purposes of this paper, *physics* simply means the science of whatever properties are needed for a complete set of laws covering such effects as stones falling, darts hitting boards, etc.

What if mental categories themselves turn out to be needed for a complete physics? What if it turns out that certain effects like darts hitting boards, etc., are not entirely determined by non-mental antecedents after all, but require essential mention of mental antecedents for their explanation? What, indeed, if telepathy, or psychokinesis, turns out to be a brute cause of certain non-mental results? I concede that we cannot exclude these possibilities absolutely, and I agree that if they obtain then the mental won't supervene on the non-mental. Yet they seem unlikely. It is one thing to hold that the categories of current physics are going to be superseded. It is quite another to hold that they are going to be superseded *inter alia* by mental categories.

In any case, even if mental categories are needed for a complete physics, the more general moral of this paper will still stand. Namely, that any categories that aren't themselves essential to the formulation of a complete physics will supervene on those which are.

University of Cambridge,
Cambridge, CB2 3RH.

REFERENCES

[1] Christopher Peacocke, *Holistic Explanation* (Oxford: Clarendon Press, 1979).

[2] Donald Davidson, 'Mental Events', in *Experience and Theory*, edited by L. Foster and J. Swanson (London: Duckworth, 1970).

[3] Jaegwon Kim, 'Causation, nomic subsumption, and the concept of event', *Journal of Philosophy* 70 (1973) 217-36.

[4] D. H. Mellor, 'The Singularly Affecting Facts of Causation', in *Metaphysics and Morality*, edited by P. Pettit, R. Sylvan and J. Norman (Oxford: Basil Blackwell, 1987).
[5] David Lewis, 'An Argument for the Identity Theory', *Journal of Philosophy* 63 (1966) 17–25.
[6] Peter Smith and O. R. Jones, *The Philosophy of Mind* (Cambridge: Cambridge University Press, 1986).
[7] Colin McGinn, *The Character of Mind* (Oxford: Oxford University Press, 1982).
[8] David Papineau, 'Social Facts and Psychological Facts', in *Popper and the Human Sciences*, edited by G. Currie and A. Musgrave (Dordrecht: Nijhoff, 1985).
[9] Jaegwon Kim, 'Supervenience and Supervenient Causation', *The Southern Journal of Philosophy* 22 (1984) 45–56.

PARADOXICAL TASKS

By CHRISTOPHER RAY

J. F. THOMSON asks us to image a lamp which may be switched on and off an infinite number of times in a finite time. If we suppose for a moment that it is not physically impossible for an infinite number of such tasks to be performed in a finite time, we may still ask with Thomson whether or not it is logically possible. We may easily imagine Thomson's reading lamp with its switch in the off position and him switching it on then off then on and so on. If it is switched on at time zero and off after one minute and on again after another 30 seconds and off again after a further 15 seconds and so on, then we might think that after two minutes we would have completed an infinite number of switching operations. But Thomson asks:

> at the end of the two minutes, is the lamp on or off? ... It cannot be on, because I did not ever turn it on without at once turning it off. It cannot be off, because I did in the first place turn it on, and thereafter I never turned it off without at once turning it on ([3], p. 5).

And this, Thomson tells us, is contradictory. He concludes that we could not in principle carry out such a super-task.

Following Benaceraff [1], Mark Sainsbury says in *Paradoxes* that Thomson's conclusion is unwarranted. He begins by distinguishing between those moments when the switching tasks are being performed (the T-series) and that first moment *after* the super-task has been completed (T*). He then argues that:

> for any moment in the T-series, if the lamp is on at that time there is a later moment in the series at which the lamp is off; and vice versa. However, nothing follows from this about whether the lamp is on or off at T*, for T* does not belong to the T-series ([2], p. 15).

The T-series is closed at one end (time zero) and open at the other 'end'. This means that although we can identify a first task at time zero, we cannot give any sense to the notion of a last task – the openness of the T-series guarantees the possibility of the tasks continuing an infinite number of times. The time T* does not occupy any point at this open end and is therefore independent of the T-series. But since T* is independent of the T-series, Sainsbury points out that our

> specification of the task speaks only to members of the T-series, and this has no consequences, let alone contradictory consequences, for how things are at T*, which lies outside the series ([2], p. 15).

So Sainsbury concludes that Thomson fails to demonstrate that the idea of a super-task is logically absurd.

Two clear implications of Sainsbury's argument are:

(1) that the lamp is either on or off at T*; and,
(2) that we cannot say either at the beginning of the super-task or once it is under way just how things will turn out at T*.

Hence there is no way to predict the state of the lamp at T*, whatever its state at the outset.

Suppose we ask an operator to carry out the super-task twice in succession in exactly the same way each time. Then there is no reason to suppose, on Sainsbury's view, that the lamp would be in the same state at each of the moments T* after the two tasks are completed. Otherwise we would always be able to predict the final state of the lamp. We may sharpen this problem as follows. Imagine now two lamps and one operator for each lamp. We ask both operators to attempt Thomson's super-task at the same time. Both lamps are off and at time zero both operators switch their lamps on. After one minute the operators switch the lamps off; after 30 seconds both lamps are switched on together; after 15 more seconds the lamps are off again; and so on. If we grant the point that T* lies outside the series, we may also grant that at T* each lamp will be either on or off. But are we forced to conclude that both lamps will be in the same state? Given that the operators begin together and continue together with the lamps flashing on and off in unison, we might expect them to 'finish' together with the lamps in precisely the same state. However, what happens during the T-series is, we are told, independent of what is the case at T*. As Sainsbury claims, *nothing* concerning T* follows from our specification of the task to be carried out because this specification relates only to times within the T-series. The fact that the lamps are initially in the same state is irrelevant, since the moment at the start of the super-task lies within the T-series. And the fact that the operators continue together does not help, because our instructions to them relate solely to times within the T-series. So why should we expect the lamps to be in the same state at T*? We are left with the unsatisfactory conclusion that two people always in step during an infinite sequence of tasks may be out of step when the super-task has been completed.

Such a conclusion seems to involve us in rather more than an *empirical* puzzle about the way things will turn out with such 'parallel' super-tasks. There seems to be a rational if not a logical inconsistency. If the super-tasks run in parallel and in step at all times during the series, then we have no reason at all to suppose that this pattern could be broken when the super-tasks are over. By accepting Sainsbury's distinction between the T-series and the (independent) time T*, then there is also now a reason to suppose that the pattern may be broken at T* when the tasks are over. The two lamps may be in different states at T* because there is no connection between the states of the lamps during the switching operations and their states at T*. Given this lack of connection, the chances of the lamps being in the identical states at T* are the same as the chances of them being in different states at T*. And this seems to be an acceptable reason for believing that the pattern may be broken. Hence, acceptance of the independence of T* from the T-series leads us to a direct conflict with our apparently reasonable expectation that two 'parallel' operations should always remain in step not just up to the end of a task but also when the task is over.

The difficulty involved in this problem seems to derive from the fact that there is no limiting behaviour of the situation to which we might appeal in order to dissolve the puzzle. When Zeno's athlete attempts to run from A to B at a uniform speed – first passing through the half-way point, then the three-quarter way point, and so on – the mathematical function which describes the athlete's progress can be represented on a graph of distance against time as a straight line. The natural limit of the line may be unambiguously defined as point B. Again, there is a natural limit for a lightly damped simple pendulum; in this case the graph of displacement from equilibrium against time shows that the amplitudes of the oscillations decrease towards the limiting value of zero displacement from equilibrium. But the lamp system has no such natural limit. For there is no preferred way for us to extend any mathematical function, which we employ to describe the behaviour of the system, across the open end of the T-series to T* itself.

It might be thought that this lack of limiting behaviour presents us with a resolution to any worry we might have concerning the parallel super-task. For it seems that we now have an explanation of the fact that two lamps which start out together can end up out of step. The mathematical function describing the system simply fails to determine which of two possible states each lamp will be in at time T*. And why should this be so problematic? There are many, many situations in physics where we cannot uniquely determine outcomes.

However, the parallel super-task raises more difficulties than that of mere uncertainty as to the final outcome. Indeed, the source of any worry we might have is the conviction that both lamps should end up in the same state, whatever that might be. Of course, we have a degree of uncertainty as to how things will turn out. Yet the uncertainty is limited to that which we would have about one lamp — we just cannot say whether the final state will be on or off. Nevertheless, as Sainsbury reminds us, we *do* know that it will be in one of these states. But we *also* have compelling reasons for believing that the lamps will be in the same state. At least we would make such claims if we believed that such a super-task is not to be ruled out on other grounds.

If we wish to deny the genuinely paradoxical nature of such super-tasks, then one of two options seems to be open to us. We may be forced to challenge our rational expectations concerning the outcomes of such tasks. Or we may maintain that only those paradoxes involving logical contradictions are genuine paradoxes. But, as noted above, we have no reason at all to take the first option. And, together with Sainsbury, we have plenty of reasons to resist the second option — in his introduction to *Paradoxes*, a paradox is described as 'an apparently unacceptable conclusion derived by apparently acceptable reasoning from apparently acceptable premisses'. And despite its lack of any outright logical contradiction, the super-task certainly falls into this category. So we should follow Thomson's advice and rule out such tasks in principle. Perhaps the moral of this tale is that if we conjure up empirical fictions, then we should not be too surprised when our stories end unhappily.[1]

Portland State University,
Portland, OR 97207-0751, U.S.A.

REFERENCES

[1] Paul Benacerraf, 'Tasks, Super-tasks, and the Modern Eleatics', *Journal of Philosophy* 59 (1962) 765–84.
[2] R. M. Sainsbury, *Paradoxes* (Cambridge: Cambridge University Press, 1988).
[3] J. F. Thomson, 'Tasks and Super-tasks', *Analysis* 15 (1954) 1–13.

---

[1] I would like to thank John Greenlees, Carol Ray, Don Moor and Peter Smith for their helpful comments on earlier drafts of this note.

DOUBLE CONDITIONALS

By ADAM MORTON

MANY if-sentences have two antecedents. Some have one antecedent nested in another, apparently of the form *if p then if q then r.* For example 'if James had come to the party then if Mary had been drunk there would have been trouble' or 'if the pound sinks any further then if interest rates do not rise then we will have serious inflation'. And others have two conjoined antecedents before a consequent, apparently of the form *if (p & q) then r.* For example 'if James had come to the party and Mary had been drunk there would have been trouble'.

I am concerned with a class of two-antecedent conditionals which are often expressed with the 'if p then if q then r' syntax, and sometimes with the 'if (p&q) then r' syntax. All the examples just given could, given suitable contexts, express conditionals of this class. In spoken English they are often expressed as 'if p and then if q then r', as in 'if James had come to the party and then if Mary had been drunk there would have been trouble'. I shall call them double conditionals, and write them *if p/q then r.* The *if* involved seems clearly to be of the family of counterfactual or subjunctive conditionals. (In fact, they are all of V. H. Dudman's third class of conditionals, those using the conditional form to ascribe a disposition or potentiality to an object or system of objects.) But — this is the central claim — if *if* is taken in accordance with the now-dominant analysis of conditionals, then double conditionals cannot be interpreted either as 'if p then (if q then r)' nor as 'if (p&q) then r'.

My orthodoxy about conditionals is the Lewis–Stalnaker possible worlds analysis as modified by Jonathan Bennett. That is, 'if p then q' is true in world w iff q is true in the nearest p-world to w. (A p-world is just a world in which p is true.) Bennett's point is that it is often not overall nearness or similarity of worlds that is needed, but nearness relative to the antecedent p. In his examples the relevant fact about the antecedent is the time it indicates, so that for example 'if there had been a nuclear war in 1980 then few would have survived' is true because the nearest worlds to actuality up to the end of 1980 which have nuclear wars continue with very few survivors, although in their later history they are very unlike actuality (for a decade, at any rate). The time indicated by the antecedent is only one relevant factor, though, so a more exact definition would therefore be: 'if p then q' is true iff q is true in all p-nearest p-worlds to w'.

Given just this machinery, we can see how double conditionals differ from various embedded conditionals. Taking *if* as just defined (and abbreviating 'p-nearest to actuality' as just 'p-nearest') three possibilities are:

(A) Embedded *if:* 'if p then (if q then r)' is true when the q-worlds q-nearest to the p-nearest p-worlds are r-worlds.
(B) Embedded *and:* 'if (p&q) then r' is true when the (p&q)-nearest (p&q)-worlds are r-worlds.
(C) Double *if:* 'if p/q then r' is true when the (p&q)-worlds q-nearest to the nearest p-worlds are r-worlds.

The difference between double *if* and embedded *if* lies in the weight given to the first antecedent, p. 'If p then (if q then r)' says: go to the nearest p-worlds and then find the nearest q-worlds to them and see if r holds. But 'if p/q then r' says: go to the nearest p-worlds and then find the nearest q-worlds which are still p-worlds and see if r holds. So in the double conditional the antecedent p still holds in the worlds in which q and r hold.

Here is a simple example in which the double conditional is true but the embedded conditions is false. I have bought a expensive and supposedly shockproof watch and I say, 'if I drop it then if it breaks I shall want my money back'. Taken as a double conditional this is surely true: in not-too-remote situations in which I have dropped the watch, if then the watch develops a fault I will be enraged and return it for a refund. But taken as an embedded conditional it can easily be false: in many of the nearest worlds in which I drop the watch it is unbroken so that the nearest worlds to them in which it is broken are worlds in which it is not dropped at all but run over by a train or melted in a crucible. And of course in those worlds I have no grounds for asking for my money back. So it is not true that if I drop the watch it will be the case that if it were to be broken I would want a refund.

This example also shows how unnatural the sense of 'if p then (if q then r)' is. People very rarely need to say things of this kind. A slight change in the example shows how the embedded conditional can be true when the double conditional is false. This time imagine that it is a very cheap and fragile watch, and I, standing on a concrete floor, say 'if I drop it then if it continues to work I will not be surprised'. This is true taken as an embedded conditional, since from the nearest worlds in which I drop it, the nearest worlds in which it continues to work are ones in which it was never dropped and I am thus not surprised. But taken as a double conditional it is false. For from the nearest worlds in which it is dropped the nearest worlds in which it has been dropped and continues to work are very remote and unexpected ones in which for example it lands in such a way that two kinds of impact cancel one another's effects, to my surprise. (Note the implicit use of Bennett's modification of the Lewis–Stalnaker condition: the worlds in which the watch is dropped and then lands gently are dropped-nearest but not overall very near to actuality. But they are the works-nearest worlds to the dropped-nearest worlds in which the watch is both dropped and continues to work. Note also how the idiom 'if I drop it and then if it continues to work then I will not be surprised', leads one on to a series of points of time structuring Bennett's modification.)

This example can be developed to show how the double conditional can be true while embedded *and* is false. Suppose that the nearest worlds in which 'I drop the watch and it continues to work' is true are those in which I first put my coat on the floor so that the watch when dropped lands on the coat. Then 'if (I drop the watch and it continues to work) then I will be surprised' is false. But the double conditional 'if I drop the watch and it continues to work then I will be surprised' is true, since the nearest worlds in which I drop the watch are not ones in which I first put my coat on the floor. (And generally, the (p&q)-worlds q-nearest to the p-nearest p-worlds need not be the same as the (p&q)-nearest (p&q)-worlds.)

The example can also be developed to show how the double conditional can be true while another embedding, 'if p then $(q \supset r)$', is false. For there might be a very far-out possibility in which the watch's dropping and breaking would lead to my surprise. Perhaps after I drop it a bat swoops down and gently rescues it before it hits the ground, but it breaks anyway. So then 'if I drop the watch then (it breaks $\supset$ I am surprised)' is false. But the double conditional is still true, since the bat-swooping world is not the nearest to the dropping world in which the watch breaks.

Similar examples can show that 'if p/q then r' can be false while 'if (p&q) then r' is true, and while 'if p then $(q \supset r)$' is true.

All these examples should work without the semantical props. That is, it should be possible to dispense with talk of possible worlds and just by giving the examples to show that in English *if* can take a meaning making 'if p then if q then r' (and 'if p and q then r', and 'if p and then if q, then r') not an instance of 'if A then B'. In fact, if you take the props away some of the examples sway a bit. The reason is that *if* can take so many other meanings besides the Stalnaker–Lewis counterfactual and the double conditional. So one thing the props do is to substitute for the effect of a full linguistic context, which narrows the range of senses a conditional can have. (And this is potentially a test for a semantical theory. Can it synthesize the expressive power of the language's use of context? It is not at all obvious whether possible worlds semantics passes this test.)

So what? One reason for being interested in double conditionals comes from issues about laws of nature. Take it to be a fairly superficial law of nature that objects released near the surface of the earth accelerate downwards at 32 feet per second per second. The law is superficial because if the history of the earth had been slightly different then things wouldn't fall at that rate but at another. The earth could easily have been a little less massive — for example if more water had evaporated into space in the past few million years or if a tiny perturbation during the formation of the earth from a disk of matter surrounding the sun had resulted in just a little less iron in the core. The objects released near its surface would accelerate downwards at a lesser rate. So the 32 feet per second law is just an amalgam of an accidental fact and a more fundamental law. Perhaps the more fundamental law is Newton's law. But that is at best an approximation to a more general truth expressed by the general theory of relativity. And that itself contains a gravitational constant whose value seems rather arbitrary. There are conjectures that its value may change with time or be determined during the early life of the universe by more fundamental factors. So we have laws beneath laws, potentially reaching down beyond our ability to make sense of them.

There are two connections with double conditionals here. One arises when we try to specify what things would be like if the laws of nature were different. For example, 'if falling things accelerated downwards at 35 ft per second per second then if there were flying machines they would need more powerful engines'. Suppose this is true. To suppose this is clearly to suppose that it is a double rather than an embedded conditional. For it could be that in the nearest situations in which gravity is just that much stronger flying machines would never have been invented. In fact, the nearest worlds to those greater-gravity worlds in which there are flying machines may be worlds in which gravity is just as it is in our world, so that more powerful engines are *not* needed. yet taken as a double conditional the assertion is true, as long as in all greater-gravity worlds flying machines can only be had, without tinkering with the force of gravity, by using more powerful engines. The conclusion is that if we wish to talk about more and less fundamental laws of nature we will need to pay attention to the kinds of conditional we use.

The other connection with double conditionals lies in the failure of an attractive idea. If we want to define 'law of nature' it is tempting to identify a law with the set of conditionals that, given the contingent facts in different worlds, it entails. Then, to make a law a proposition like any other it would have to be the case that the set of possible worlds in which a law holds is identical to the set of worlds in which a set of embedded *if*s hold, those of the form 'if F then if q then r' where F is the fact about a particular world that 'triggers' the conditional 'if q then r'. (E.g. F might be 'The earth has mass m' and 'if p then q' might be 'if *o* is dropped it will accelerate downwards at 32 ft/sec/sec'.) For a very superficial law like the 32 ft/sec/sec law the F will be redundant, and for a very profound law the r will itself be a conditional. The interesting point is that this idea does not work.

It does not work because a conditional can fail to be true in a world although the world is subject to the relevant law. And the reason for this is the same as the reason that the double conditional is not the same as an embedded conditional. Thus it is a law of our world that objects dropped near the surface of the earth accelerate towards the earth at 32 feet per second per second, but some conditionals of the form 'if *o* were dropped it would fall at 32 ft/sec/sec' are false because the nearest world in which *o* is dropped is one in which the earth has a different gravity. (And, similarly, a law can apply to a world even though the corresponding conditionals are false in it.) So a law is not a proposition true in all worlds in which a set of conditionals is true.

That is a purely negative result, though an interesting one. (Perhaps it is one of the reasons that David Lewis, when giving an account of laws in terms of possible worlds, takes a completely different line.) But there may be a positive result lurking here. The counterexamples show that the conditionals 'if p then q' that it is natural to associate with a law of nature L are not all true, taken as isolated subjunctive conditionals. But the associated double conditionals 'if L/p then q' *are* true. This suggests a criterion of lawfulness: a law of nature is a proposition L associated with a set of conditionals S, such that the set of worlds in which all of the double conditionals 'if L/p then q' are true (for all 'if p then q' in S) meets two conditions. (i) It includes the actual world and (ii) it is identical to the set of worlds in which L is true.

Standard examples of laws of nature apparently pass this test, and traditional examples of accidental generalizations, such as 'all the coins in my pocket are copper', apparently do not. So we have an interesting open question: are there plausible examples of laws or non-laws which, given reasonable intuitions about the structure of possible worlds, provide counterexamples to this conjecture?[1]

University of Bristol,
9 Woodland Road, Bristol BS2 8XY

---

[1] This paper is an expansion of a fragment from a much larger draft by Fabrizio Mondadori and me. I am grateful for Mondadori's advice (and some of the ideas are his). The Editor of *Analysis* found, again, a serious mistake in an earlier draft. For the orthodoxy about counterfactional conditionals see David Lewis, *Counterfactuals* (Oxford: Blackwell, 1977), and Jonathan Bennett's 'Counterfactuals and Possible Worlds', *Canadian Journal of Philosophy* 5 (1974) 381–402. For Lewis' treatment of laws of nature see *Counterfactuals* and also his *On the Plurality of Worlds* (Oxford: Oxford University Press, 1987). Among the many papers in *Analysis* on conditionals the following are particularly relevant: Michael Clarke 'Ifs and Hooks' *Analysis* 34 (1973) 72–83; J. S. Edwards 'A confusion about "if ... then"' *Analysis* 34 (1973) 84–90; A. J. Dale, 'A defence of material implication', *Analysis* 34 (1973) 91–5; and V. H. Dudman, 'Parsing "if" sentences', *Analysis* 47 (1987) 193–9. Dale argues that 'if p then (if q then r)' is equivalent to 'if (p and q) then r'. Since the 'shockproof watch' example above is a counterexample to this, it might be reasonable to suppose that Dale's claim holds for at most Dudman's first category of conditionals.

Some related points about embedded conditionals are found in Adam Morton 'Would cause', *Proc. Aristotelian Soc.* 81 (1980/1) 139–51. A more complete working out of the suggestion about laws of nature made at the end of the paper would have to incorporate the related points made in Adam Morton 'If I were a dry well-made match', *Dialogue* 12 (1973) 322–4, and Peter van Inwagen 'Laws and Counterfactuals', *Noûs* 13 (1979) 439–453.

CONDITIONALS, CONTEXT, AND TRANSITIVITY

By E. J. LOWE

SUPPOSE that two people are discussing the influence of upbringing and social background upon a person's political convictions, and one of them, *X*, takes Margaret Thatcher as an example of someone who, though a firm supporter of the capitalist free market economy, might have had a quite different attitude towards it. *X* reasons as follows:

> (A) If Margaret Thatcher had been born and brought up in the Soviet Union, she would have had communist sympathies; and if she had had communist sympathies, she would have been opposed to the capitalist free market economy. So, if she had been born and brought up in the Soviet Union, she would have been opposed to the capitalist free market economy.

X's reasoning seems unexceptionable to the point of appearing almost banal. Next suppose that two other people are discussing what sorts of political orientation are necessary for success in British politics, and one of them, *Y*, takes Margaret Thatcher as an example of someone who, though she has risen to the peak of political power in Britain, might have fared quite differently in her career here. *Y* reasons as follows:

> (B) If Margaret Thatcher had had communist sympathies, she would have joined the British Communist Party; and if she had joined the British Communist Party, she would not have become Prime Minister of the U.K. So, if she had had communist sympathies, she would not have become Prime Minister of the U.K.

Again, *Y*'s reasoning seems almost banal in its uncontentiousness. Finally, however, suppose that a third person, *Z*, overhears parts of each conversation (perhaps on crossed telephone lines), mistakenly thinking that he is listening to part of a single conversation, and that what he hears is this:

> (C) If Margaret Thatcher had been born and brought up in the Soviet Union, she would have had communist sympathies; and if she *had had communist sympathies, she would have joined the British Communist Party.* So, if she had been born and brought up in the Soviet Union, she would have *joined the British Communist Party.*

Here the unitalicized part comes from *X*'s conversation and the italicized part from *Y*'s conversation.

Now, the piece of reasoning (C) which *Z* mistakenly thinks is being voiced is *formally* exactly like the two uncontentious pieces of reasoning (A) and (B) voiced by *X* and *Y* respectively. Moreover, one might suppose that *Z* could even allow that this piece of reasoning has true premisses: for the first and second premisses of (C) are, respectively, (A)'s first premiss and (B)'s first premiss. Yet the conclusion of (C) seems patently false. And indeed inferences like (C) are standardly offered as illustrating the *non-transitivity* of subjunctive conditionals in natural language (see, e.g., Stalnaker [6]). But in view of the uncontroversial character of inferences (A) and (B), and in the light of the way in which we have constructed (C) from snatches of conversation wrenched out of their original contexts, it seems more plausible to suggest that what is wrong with (C) has nothing to do with the *form* of the inference employed, but rather something to do with the difficulty of finding an intelligible conversational *context* for the co-assertion of the premisses. The first premiss of (C) is perfectly acceptable within the context envisaged for the conversation involving *X*, and the second premiss of (C) is perfectly acceptable within the context envisaged for the conversation involving *Y*; but it is apparently impossible to find any *single* conversational context in which ***both*** premisses of (C) would be acceptable. Moreover, connected with this fact is a strong intuition that the fallacy involved in (C) is some sort of fallacy of *equivocation*: one senses that the consequent of (C)'s first premiss and the antecedent of its second premiss, though virtually identical at a lexical level, do not match in content. An explanation of this intuition is something which any adequate account of the semantics of natural-language conditionals should provide. We may thus expect such an account to hold that what examples like (C) go to show is *not* that natural-language subjunctive conditionals are non-transitive but rather that they need to be assessed for truth-value relative to a context (with the implication that their content is context-dependent), and that they do in fact sustain transitivity within context. (Cf. Wright [9] and [10]. Although Wright supports these claims, he does not back them up with very much in the way of a semantic theory of subjunctive conditionals. I shall, incidentally, continue to talk in this way about 'subjunctive conditionals' as though these constituted a well-defined and unitary logico-semantic category, despite the recent well-aimed strictures of Bennett [1]; this larger issue is too large for my present brief.)

Contemporary orthodoxy concerning the semantics of natural-language subjunctive conditionals, embodied in the Stalnaker–Lewis (SL) approach, cannot be exactly correct if the foregoing line of thought is right. According to the SL approach, we may say (to a first approximation which suffices for my purposes):

T1 $p \mathrel{\Box\!\!\rightarrow} q$ is true in a world *w* iff either there is no p-world or else q is true in the p-world most similar to *w* where by a 'p-world' is meant quite simply a world in which p is true. Here it should be remarked that assessing degrees of *overall* similarity between worlds (which are what are relevant to the truth-conditions of subjunctive conditionals according to T1) involves making judgements about the relative weights with which similarities in particular *respects* contribute to overall similarity, and the SL approach can readily allow that different weighting functions may be induced by different conversational contexts (cf. van Fraassen [8], pp. 193 f.). (On the SL approach, this will be a matter for the *pragmatics* rather than for the *semantics* of subjunctive conditionals.) For instance, the SL approach may allow that in one context the weightings are such that similarity in respect of an individual's place of birth and upbringing in different worlds contributes more strongly to overall similarity between those worlds than does similarity in respect of the influence which social background has upon an individual's political views, whereas in another context these weightings may be reversed. However, it is evident from T1 that the SL approach must regard subjunctive conditionals as non-transitive, and hence the following inference-pattern as invalid:

$$\text{P1}\quad \begin{array}{c} p \mathbin{\Box\!\!\rightarrow} q \\ q \mathbin{\Box\!\!\rightarrow} r \\ \hline p \mathbin{\Box\!\!\rightarrow} r \end{array}$$

Nothing that the SL approach may say about the context-dependency of similarity-weightings can affect this point. And hence the SL approach just has to represent inferences like (A) and (B), quite as well as ones like (C), as being straightforwardly invalid. Yet this, as we have seen, is counterintuitive.

The best that the SL approach can do by way of retrieving this situation is to suggest that inferences like (A) and (B) are tacitly or implicitly inferences of the form

$$\text{P2}\quad \begin{array}{c} p \mathbin{\Box\!\!\rightarrow} q \\ (p \mathbin{\&} q) \mathbin{\Box\!\!\rightarrow} r \\ \hline p \mathbin{\Box\!\!\rightarrow} r \end{array}$$

which is perfectly valid according to the SL approach. (For such a suggestion, see, e.g., Jackson [2], p. 82.) Now, of course, it may well be that whenever in ordinary language we seem to reason in accordance with inference-pattern P1, we will be willing to strengthen our second premiss so as to make our reasoning conform to inference-pattern P2. But this does not by any means show that we are 'really' tacitly employing P2 rather than P1: for it may simply be that (as I believe) the correct logic of conditionals implies that whenever two conditionals $p \mathbin{\Box\!\!\rightarrow} q$ and $q \mathbin{\Box\!\!\rightarrow} r$ are jointly held true, so too should be the conditional $(p \mathbin{\&} q) \mathbin{\Box\!\!\rightarrow} r$. So the fact that we would indeed be willing to strengthen (A) and (B) — but not (C) — in this way does not constitute evidence favouring the SL approach. Indeed, when we focus on (C) rather than on (A) and (B), we can see that the considerations now being raised in fact count *against* the SL approach. For if, as the SL approach implies, P1 were straightforwardly invalid and yet P2 valid, then it should suffice to explain our rejection of (C) that we can readily entertain the joint truth of (C)'s premisses while dismissing as false the appropriately strengthened version of (C)'s second premiss (i.e., 'If Margaret Thatcher had been born and brought up in the Soviet Union and had had communist sympathies, she would have joined the British Communist Party'). But in fact we have *severe difficulty* in entertaining the joint truth of (C)'s original premisses (that is, in thinking of a context in which we would be prepared to assert them together), and this phenomenon is evidently closely connected with our sense that (C) involves some sort of equivocation. In short, the diagnosis which our rejection of (C) receives at the hands of the SL approach simply fails to accommodate the main locus of our unease about (C). If the diagnosis were correct, we ought to be perfectly happy to co-assert (C)'s premisses, rejecting its conclusion merely because being unwilling to strengthen its second premiss, and discerning in such co-assertion no hint of equivocation. But such is *not* our intuitive position. (See Wright [9] and [10] for confirmation of this point.)

What we need, then, is an alternative account of the semantics of natural-language subjunctive conditionals which will be more faithful to our linguistic intuitions concerning the examples presented above. For instance, we might (to a first approximation, at any rate) regard the natural-language subjunctive conditional as a species of *strict* conditional — thus securing the validity of inference by transitivity for such conditionals — while at the same time making due allowance for context dependency. (The feasibility in principle of such an alternative to his own approach is recognized by Stalnaker in his [7], pp. 125 f. Lewis [3], p. 13, rejects the idea, for reasons to which I shall return later.) What we could thus say is:

T2 $p \mathrel{\Box\!\!\rightarrow} q$ is true in a context C in a world w iff $\Box_{Cw}(p \supset q)$

where '$\Box_{Cw}$' is a necessity operator which may be read as meaning: 'In every possible world sufficiently similar to w as determined by C, it is true that . . .', where C determines both the weights with which similarities in particular respects contribute to overall similarity between worlds and also what qualifies as a sufficient level of overall similarity. (A certain amount of vagueness may infect this reading in at least some contexts, where the contextually determined factors are imprecise: but this is as it should be, since in such contexts at least some conditionals will not be assign-able a determinate truth-value.) Now, of course, the inference-pattern

$$\text{P3}\quad \begin{array}{l} \Box_{Cw}(p\supset q) \\ \Box_{Cw}(q\supset r) \\ \hline \Box_{Cw}(p\supset r) \end{array}$$

is straightforwardly valid, and so correctly (in my view) represents the inferences (A) and (B) above as being perfectly sound. But in inference (C) it appears that the context cannot intelligibly be kept constant, and accordingly that this pattern cannot be supposed to be instantiated in it (though a more considered verdict on the matter must wait a moment).

Observe, in particular, how this approach would treat the subjunctive conditional

(1) If Margaret Thatcher had had communist sympathies, she would have joined the British Communist Party

as it might be asserted within the conversational contexts envisaged for inferences (A) and (B) respectively. In the context of the first conversation, (1) should be dismissed as *false*: because in that context it is *not* the case that in every possible world 'sufficiently similar' to the actual world in which Margaret Thatcher has communist sympathies, she also joins the British Communist Party. For in this context her place of birth and upbringing carries negligible weight in global similarity-rankings, so that worlds in which she is a native Soviet citizen are indeed 'sufficiently similar' to the actual world to be taken into consideration. And in most such worlds she obviously does *not* join the British Communist Party, nor indeed any British political party whatsoever. By contrast, in the context of the second conversation, (1) may plausibly be assessed as *true*, because in this context Margaret Thatcher's place of birth and upbringing (though not the influence of that upbringing upon her political convictions) carries very considerable weight in global similarity-rankings, so that all the 'sufficiently similar' worlds in this context are worlds in which Margaret Thatcher is a British citizen (and, moreover, one bent upon a career in British politics). And in all such worlds in which she has communist sympathies, we may plausibly suppose, she *does* give rein to her political ambitions by joining the British Communist Party.

However, while T2 is in some ways an improvement on T1, it is still not entirely satisfactory because — for one thing — it renders vacuously or trivially true a subjunctive conditional $p \Box\!\!\rightarrow q$ in a context *C* and world *w* provided merely that $\Box_{Cw}\sim p$ obtains. (By contrast, T1 only renders vacuously true a subjunctive conditional with a *logically* impossible, i.e. inconsistent, antecedent — which may perhaps be deemed a harmless enough bit of legislation.)

Consider, thus, the conversational context envisaged for inference (B) above. In this context it seems clear that there is *no* world sufficiently similar to the actual world in which Margaret Thatcher is born and brought up in the Soviet Union, but this means that T2 implies that the following conditional (which is of course the conclusion of inference (C)) is *true* in such a context:

> (2) If Margaret Thatcher had been born and brought up in the Soviet Union, she would have joined the British Communist Party

and this seems very hard to accept, not least because it then emerges that according to T2 inference (C) is in fact sound (having true premisses and a true conclusion) within the context envisaged for inference (B) (though not, of course, within that envisaged for inference (A)).

There are various ways in which we might seek to modify T2 so as to overcome this problem. One would be to offer in its place

T3 $p \mathrel{\Box\!\!\rightarrow} q$ is true in a context *C* in a world *w* iff $\Box_{Cw}(p \supset q)$ & $\Diamond_{Cw} p$

Another would be to propose instead

T4 $p \mathrel{\Box\!\!\rightarrow} q$ is true in a context *C* in a world *w* iff $\Box_{Cw}(p \supset q)$ & $(\Diamond_{Cw} p \vee \Box_{Cw} q)$

Both T3 and T4 still secure transitivity, but neither delivers the verdict that (2) is true in the context envisaged for inference (B). Accordingly, both T3 and T4 have the desirable implication that inference (C) is sound *neither* in the context envisaged for inference (A) *nor* in that envisaged for inference (B). This does not prove, of course, that there is *no* conceivable context, however bizarre and unnatural, in which T3 or T4 could represent (C) as being a sound inference. Nor do I need to prove this. For in order to explain and vindicate our intuitive rejection of (C) consistently with upholding transitivity it is enough to show that T3 and T4 do not represent (C) as sound in the sorts of context in which either of its premisses would *normally* be regarded as expressing truths. For linguistic intuitions reflect normal usage and have no legislative force beyond it.

I shall not attempt to arbitrate here between T3 and T4, nor do I wish to imply that no other options are available (but see Lowe [4] for further discussion). I do believe that when we assert a subjunctive conditional in ordinary language we normally intend to convey the fact that we regard its antecedent as expressing a 'possibility', *in the relevant contextually determined sense.* (This is so even where we may take the antecedent to express some sort of *im*possibility in another sense — for instance, when in the context of a mathematical *reductio* we assert something like 'If there were a greatest prime number, then . . .'. Even here, I believe, we are in some sense endeavouring to entertain the 'possibility' that there be a greatest prime number, albeit only in order to dismiss that thought. Cf. Lewis [3], p. 24.) But whether, as T3 implies, this is because $p \mathbin{\Box\!\!\rightarrow} q$ (in *C*, *w*) actually *entails* $\Diamond_{Cw}p$, or merely because $p \mathbin{\Box\!\!\rightarrow} q$ (in *C*, *w*) conversationally/conventionally *implicates* $\Diamond_{Cw}p$, is a difficult matter on which to adjudicate.

Finally, I should remark that the advantages of T3 and T4 over T2 are not limited to their treatment of subjunctive conditionals like (2). It is a particular strength of both T3 and T4 that, like T1 but unlike T2, they validate neither inference by contraposition nor inference by strengthening the antecedent, that is, neither of the following inference-patterns:

$$\text{P4} \qquad \frac{p \mathbin{\Box\!\!\rightarrow} q}{\sim q \mathbin{\Box\!\!\rightarrow} \sim p} \qquad\qquad \text{P5} \qquad \frac{p \mathbin{\Box\!\!\rightarrow} q}{(p \,\&\, r) \mathbin{\Box\!\!\rightarrow} q}$$

though, importantly, they do validate appropriately weakened versions of each, namely, the inference-patterns below:

$$\text{P6} \qquad \frac{\begin{array}{c} \sim \Box q \\ p \mathbin{\Box\!\!\rightarrow} q \end{array}}{\sim q \mathbin{\Box\!\!\rightarrow} \sim p} \qquad\qquad \text{P7} \qquad \frac{\begin{array}{c} \sim (p \mathbin{\Box\!\!\rightarrow} \sim r) \\ p \mathbin{\Box\!\!\rightarrow} q \end{array}}{(p \,\&\, r) \mathbin{\Box\!\!\rightarrow} q}$$

It is significant that the rejected stronger versions are both forms of inference which have much less intuitive appeal than inference by transitivity, so far as natural-language subjunctive conditionals are concerned, so that T3 and T4 occupy an attractive area of middle ground between T1 and T2. (It is true that it is commonly supposed that no such area of middle ground can exist, because it is assumed that acceptance of transitivity commits one to acceptance of strengthening the antecedent, via the assumption that $(p \,\&\, q) \mathbin{\Box\!\!\rightarrow} p$ is a logical truth: see, e.g., Stalnaker [6]. But this latter assumption is questionable, for the intuitions which seem to support it can be explained by T3 and T4 as arising rather from the fact that $\Diamond_{Cw}(p \,\&\, q)$ entails $(p \,\&\, q) \mathbin{\Box\!\!\rightarrow} p$ (in *C*, *w*) together with the fact that since conditionals of this form never feature in ordinary language reasoning, no natural conversational context for their assertion suggests itself: for this latter fact encourages us to interpret them relative to the 'null' context, in which $\Diamond_{Cw}(p \,\&\, q)$ expresses the bare logical possibility or consistency of the conjunction, so that $(p \,\&\, q) \mathbin{\Box\!\!\rightarrow} p$ comes to be regarded as uncontentiously true at least unless its antecedent is inconsistent. See further Lowe [5]. Here it is appropriate to note that the argument in Lewis [3], pp. 10 ff., against treating subjunctive conditionals as a species of strict conditional turns on the validity for strict conditionals of strengthening the antecedent, and so may count against T2 but not against T3 or T4.) One last point: T3 and T4, like T2 but unlike T1, have the considerable attraction of permitting the reduction of the logic of subjunctive conditionals to a branch of monadic modal logic, with all the gains of economy and simplicity which such a reduction brings (see further Lowe [4]). We have to hand, then, an approach to the semantics of natural-language subjunctive conditionals which at once respects our linguistic intuitions better than does the SL approach and offers the prospect of a simpler conditional logic. Why not buy it?

University of Durham,
50 Old Elvet, Durham DH1 3HN.

REFERENCES

[1] Jonathan Bennett, 'Farewell to the Phlogiston Theory of Conditionals', *Mind* 97 (1988) 509–27.
[2] Frank Jackson, *Conditionals* (Oxford: Basil Blackwell, 1987).
[3] David K. Lewis, *Counterfactuals* (Oxford: Basil Blackwell, 1973).
[4] E. J. Lowe, 'A Simplification of the Logic of Conditionals', *Notre Dame Journal of Formal Logic* 24 (1983) 357–66.
[5] E. J. Lowe, '"If A and B, then A"', *Analysis* 45.2 (March 1985) 93–8.
[6] Robert C. Stalnaker, 'A Theory of Conditionals', in *Studies in Logical Theory*, edited by N. Rescher (Oxford: Basil Blackwell, 1968).
[7] Robert C. Stalnaker, *Inquiry* (Cambridge, Mass.: MIT Press, 1984).
[8] Bas C. van Fraassen, 'Essences and Laws of Nature', in *Reduction, Time and Reality*, edited by R. Healey (Cambridge: Cambridge University Press, 1981).
[9] Crispin Wright, 'Keeping Track of Nozick', *Analysis* 43.3 (June 1983) 134–40.
[10] Crispin Wright, 'Comment on Lowe', *Analysis* 44.4 (October 1984) 183–5.

WHY IS BELIEF INVOLUNTARY?

By JONATHAN BENNETT

THIS paper will present a negative result — an account of my failure to explain why belief is involuntary. When I announced my question a year or so ahead of time, I had a vague idea of how it might be answered, but I cannot make it work out. Necessity, this time, has not given birth to invention.

Still, my tussle with the question may contribute either towards getting it answered or showing that it cannot be answered because belief can be voluntary after all.

Most of the paper was written while I expected to get the question answered, and I have chosen not to hide that fact by revising the tone. I offer the paper as an essay in the 'analytic' manner in Descartes's sense of that term. It is the manner of the *Meditations* — a presentation in the order of discovery or, in my case, of non-discovery.

I Introduction

Could I get you to believe something by offering you an inducement to believe it? That is, could belief be produced by threats or bribes? Well, yes, in a couple of uninteresting ways; let us set them aside first.

(i) You don't think I would ever threaten anybody, and you tell me this; I respond convincingly: 'Believe that I am capable of threats! If you don't, I'll kick you on both shins.' That might get you to acquire the belief, but only by giving you evidence for it, not by inducing you to acquire it. You would have acquired it just as easily and surely if you had heard me threaten someone else. My question is: Could I get you to believe something by offering you an inducement to believe it without thereby providing evidence for it? Or, for brevity, Could I induce you to believe something?

(ii) Pascal held that one's chances of believing Christianity could be improved by keeping pious company, reading holy books, and praying. 'That will dull your mind and make you believe,' he charmingly said. Other kinds of mind-dulling can also produce credulity. We can make a person temporarily believe in UFOs or ESP (for example) by perceptually isolating him for several days and then barraging him with propaganda, and beliefs can be caused by post-hypnotic suggestion. There may be other such techniques, and some of them may be quite reliable. Clearly, I could induce you to start up one of those procedures leading to your acquiring a specified belief. My question is: Could I *immediately* induce you to believe something? That is, could an inducement offered by me bring it about that you believe something, where (i) I don't provide evidence for the truth of the belief and (ii) this does not work by your being induced to do something that leads to your acquiring the belief?

I immediately induce you to raise your arm if I say 'If you raise your right arm I'll make it worth your while' and you, greedy or curious, immediately raise the arm. This is an immediate inducement because you are not induced to do something that has as a consequence your raising your right arm. (There are indeed intermediate steps: you come to believe that I have offered the inducement, and you form an intention to raise your arm. But you are not induced to do either of those — they are constitutive parts of your being induced to raise your arm, not means you are induced to take to getting your arm up.) The inducement would be mediated if it induced you to arrange for someone else to bully you into raising your arm, or to get into a situation where you would have a nasty accident if you didn't raise it, or to pull your right arm up with your left hand. I immediately induce you to do A if your response is to do A 'just like that', as Bernard Williams has said.

It will be important later on, but I had better say it now: What one is induced to do — or intends or plans or wants or aims or tries to do — is to bring about a state of affairs. When Norman hit the ball in my direction, I caught it, thereby producing a particular catch; but as I ran and extended my hand I was not trying to produce *that catch*, but rather to produce *a catch* or (better still) to make it the case that *I caught the ball*. Quite generally, inducements should be thought of as inducements to *make it the case that P* for various values of P. This can be defended by argument,[1] but in this paper I offer it as an axiom of the discussion.

A possible trap is worth a paragraph. When I offer you a reward if you will raise your arm and you raise it 'just like that', you don't set yourself to do something which will lead to your arm's going up; so the offer of a reward does not relate in a *motivationally mediated* way to your arm's going up. Still, your arm wouldn't go up unless certain neural events occurred first; so the offer of a reward relates in an *ontologically mediated* way to your arm's going up — the offer and the gesture lie on a causal chain with other items in between. It is clear that although the neural events mediate in the causal sequence, no motivating thought about them intervenes in the practical syllogism that takes you from the reward premiss to the action conclusion. If one doesn't distinguish the two kinds of (im)mediacy one will think that belief is on a par with behaviour as regards voluntariness: each can be induced, and neither can be induced immediately.[2] But this masks what may be a real difference, namely that between what is motivationally (and thus also ontologically) mediated and what is merely ontologically mediated. It may be the case that behaviour can and belief cannot be induced in a motivationally immediate manner.

That is the question. Can I immediately induce you to acquire a belief? Can I, without giving you evidence for the belief and just by offering convincing bribes or threats, get you to acquire the belief 'just like that', immediately, without setting your thought on producing some intervening event that will lead to your getting the belief? The inducement need not be offered by another person — an impersonal part of the world would do as well. What matters is whether one can acquire a belief because one sees an advantage in having it, this being done as immediately as in raising one's arm because one thinks it will be good for it to be up.

It is widely held that no-one could do this, and I think the usual view is that the impossibility is strong to the point of being conceptual or logical. Most of those who think about the question have found it incoherent or absurd to suppose anyone should acquire a belief, just like that, *simply* because he wanted to, as though acquiring a belief were like raising one's arm. I agree. There is indeed something so chokingly unswallowable about the idea of someone's voluntarily coming to believe something that I have to suspect that this is ruled out at a deeper level than the contingent powers of our minds.

---

[1] It is obliquely defended in J. Bennett, *Events and their Names* (Indianapolis: Hackett, 1988), ch. 14.

[2] Thus Margery Bedford Naylor, 'Voluntary Belief', *Philosophy and Phenomenological Research* 45 (1985) 427–436, especially p. 430.

I want to get beyond the level of mere hunch or intuition, by explaining why voluntary belief is conceptually impossible.

(When I say that belief cannot be voluntary, I mean that nobody could be immediately induced to believe something. For James Montmarquet, the involuntariness of belief consists in the fact that what one can believe is normally constrained by what theoretical reasons one has, which he compares with the fact that how one can act is constrained by what practical reasons one has. He infers that believing and acting are on a par so far as voluntariness is concerned.[3] This is indeed a respect in which belief and behaviour are alike, but it does not make them both involuntary. Anyone who thinks it does must be taking voluntariness to require freedom from all constraints, total unconditionedness. In company with Locke, Leibniz, Hume and many others, I understand voluntariness as responsiveness to practical reasons. I take ϕing to be voluntary if one's ϕing depends upon inducements, that is, if one can ϕ or not depending on whether one thinks one has *practical* reasons to ϕ. Actions are voluntary in that sense, and beliefs seem not to be.)

II ACCEPTANCE VS. SUBJECTIVE PROBABILITY

I am using 'believe that P' to mean 'regard P as highly likely'. My topic, then, is the impossibility of voluntarily raising the probability that one accords to a proposition. That is not the same as voluntarily *accepting* a proposition.[4]

It is not unreasonable to use 'belief' to stand for acceptance. The word is so used by those who have most strongly emphasized the difference between acceptance and high subjective probability. Here is a reason for using the term their way. Suppose the contrary: belief is a matter of high subjective probability; for example, to believe that P is to give it at least a 0.95 chance of being true. Then consider a fair lottery with a thousand entrants. What is your subjective probability for 'This entrant will not win the lottery' as applied to each entrant? It had better be 0.999. So,

[3] James Montmarquet, 'The Voluntariness of Belief', *Analysis* 46 (1986) 49–53.

[4] For helpful elaborations of the distinction, see Ronald B. de Sousa, 'How to Give a Piece of Your Mind: or, The Logic of Belief and Assent', *Review of Metaphysics* 25 (1971) 52–79; Mark Kaplan, 'Rational Acceptance', *Philosophical Studies* 40 (1981) 129–145; Patrick Maher, 'The Irrelevance of Belief to Rational Action', *Erkenntnis* 24 (1986) 363–384. I am indebted to Patrik Maher for my knowledge of the other two papers.

according to the present view, you *believe*, of each entrant, that he or she won't win; you also believe that these are all the entrants; yet it would be unreasonable for you to believe that no-one will win. This seems to be a case where it is utterly reasonable to believe that P, which obviously and elementarily entails Q, yet it is utterly unreasonable to believe that Q. That is incredible. A natural response to this is the thought: 'Do I really believe that Smith won't win? Perhaps I don't. I think he has a vanishingly small chance of winning, but his not winning isn't something I accept, adopt, make my own, stand behind.' That is the thought that acceptance should be distinguished from high subjective probability, and that 'believe' and its cognates should go with the former. Peter van Inwagen, who has helped me in my thinking about this, suggests that subjective probabilities are just a species of acceptance: to regard P as likely is to accept the proposition that probably P.

I am sure that it is right to distinguish acceptance from high subjective probability, and it seems right to tie the word 'belief' to the former. I shall not do so, however. The thoughts I have to offer concern subjective probabilities, and I want 'belief' as a convenient word for them. If van Inwagen is right that the two relate as genus to species, they probably stand or fall together so far as (in)voluntariness is concerned, but this is not a topic I have explored.

One's acceptance of a proposition seems to be a state or condition that one can be in, but it can be associated with one or more kinds of intellectual act — affirming P in one's heart, turning one's thoughts away from P's chances of being false, throwing one's weight behind P in one way or another.

Now, many philosophers have seemed to say that belief can be voluntary, but I don't think they have been contradicting the presupposition of the present paper. It seems fairly clear that what they have been treating as voluntary is an intellectual act — often an inner speech act — that is associated with acceptance. Newman, for example:

> Assent is an act of the mind, congenial to its nature; and it, as other acts, may be made both when it ought to be made, and when it ought not. It is a free act, a personal act for which the doer is responsible ... .[5]

Similarly, Descartes says that we have 'a kind of freedom that enables us always to refrain from believing things which are not completely certain', which might refer to acceptance or subjective probability; but he introduces that passage with the following: 'We have free will, enabling us to withhold our assent in doubtful matters ...', which is clearly about some intellectual act rather than about belief.[6] Chisholm writes:

---

[5] J. H. Newman, *An Essay in Aid of a Grammar of Assent* (Westminster, Md.: Christian Classics, 1973), p. 232.

> 'Belief' may ... refer to a psychological *act* — to the fact that a man *accepts* a proposition, or *takes* it to be true. In this active sense of 'belief', a belief ... may be called wrong and in much the same sense as that in which a criminal act may be called wrong.[7]

I am not sure what it is that Chisholm here treats as voluntarily and potentially culpable, but I don't think it is the assigning of subjective probabilities, which is my topic in this paper.[8]

III WILLIAMS'S FIRST REASON

Bernard Williams asserts that 'it is not a contingent fact that I cannot bring it about, just like that, that I believe something, as it is a contingent fact that I cannot bring it about, just like that, that I'm blushing'; and he offers two reasons why this is so.[9]

The first of them, Williams says, 'is connected with the characteristic of beliefs that they aim at truth. If I could acquire a belief at will,' he continues, 'I could acquire it whether it was true or not. If in full consciousness I could will to acquire a "belief" irrespective of its truth, it is unclear that before the event I could seriously think of it as a belief, i.e. as something purporting to represent reality.' If this were offered as a free-standing reason why beliefs cannot be immediately induced, it would fail. Before the event, one can think: 'I shall get myself to believe that P; the belief will be false, and I shall be deluded, but that is what I want.' This acknowledgment of falsity does not prevent me from thinking of the aimed-at state as one of belief.

Williams might agree, for he seems not really to trust this line of thought, and proceeds with a cautious variant on it:

> At the very least, there must be a restriction on what is the case after the event; since I could not then, in full consciousness, regard this as a belief of mine, i.e. something I take to be true, and also know that I acquired it at will. With regard to no belief could I know — or, if this is to be done in full consciousness, even suspect — that I had acquired it at will. But if I can acquire beliefs at will, I must know that I am able to do this; and could I know that I was capable of this feat, if with

---

[6] René Descartes, *Principles of Philosophy* 1: 6.

[7] Roderick Chisholm, 'Lewis's Ethics of Belief', in *The Philosophy of C. I. Lewis*, ed. P. A. Schilpp (La Salle, Ill.: Open Court), 223–242, at p. 223.

[8] I found the Newman and Chisholm references in Louis P. Pojman, 'Believing and Willing', *Canadian Journal of Philosphy* 15 (1985) 37–55, at p. 38. Pojman also cites Thomas Aquinas, *Summa Theologica* Part II, Q4, Art. 2; William James, 'The Will to Believe', in *Essays in Pragmatism* (New York: Hafner, 1969); Jack Meiland, 'What we Ought to Believe or the Ethics of Belief Revisited', *American Philosophical Quarterly* 17 (1980).

[9] Bernard Williams, 'Deciding to Believe', in his *Problems of the Self* (Cambridge: Cambridge University Press, 1973), 136–151, at pp. 148–151.

> regard to every feat of this kind which I had performed I necessarily had to believe that it had not taken place?

We are invited to answer No, but I answer Yes. Here is why.[10]

Credam is a community each of whose members can be immediately induced to acquire beliefs. It doesn't happen often, because they don't often think: 'I don't believe that P, but it would be good if I did'. Still, such thoughts come to them occasionally, and on some of those occasions the person succumbs to temptation and wills himself to have the desired belief. (Sometimes he merely wants to be the centre of attention and to amuse others. Someone who has no skill as an actor can instantly start to behave *exactly* as though he believed that P, by coming to believe that P. It is fun to watch it happen.) When a Credamite gets a belief in this way, he forgets that this is how he came by it. The belief is always one that he has entertained and has thought to have some evidence in its favour; though in the past he has rated the counter-evidence more highly, he could sanely have inclined the other way. When he wills himself to believe, that is what happens: he wills himself to find the other side more probable. After succeeding, he forgets that he willed himself to do it.

The community have tried inducing paid volunteers to immediately acquire beliefs that have previously been deeply and radically at odds with their belief systems, and they have always failed. The trick cannot be worked if the protective forgetfulness would require that the rest of the person's beliefs be drastically rearranged.

After successfully willing himself to have a certain belief, a Credamite may later get evidence that that is what he has done; e.g. someone may tell him. Then he either rejects the evidence (e.g. disbelieves the informant) or else accepts that he has willed himself to have the belief, and usually loses the belief because of his knowledge of how he got it.[11] In the latter case, he still remembers having had the belief, and now remembers willing himself to acquire it. The belief itself suppressed the memory of how it was caused; when the belief went, the memory emerged. So each Credamite knows that he sometimes wills himself to believe something, even though it is never true that he *now* has a belief which he *now* remembers having willed himself to acquire.

There is my case for answering Yes to Williams's question. I do not believe that my story is possible because I do not think that one could immediately induce beliefs. But the story satisfies Williams's demands; it is possible, for all he shows to the contrary;

---

[10] For a more detailed discussion of this argument of Williams's, see Barbara Winters, 'Willing to Believe', *Journal of Philosophy* 76 (1979) 243–256.

[11] Usually—but sometimes he retains the belief because since voluntarily acquiring it he has encountered evidence for it.

so his first reason does not explain why voluntary belief is impossible.

Furthermore, the supposed reason applies equally to any acquisition of belief, immediate or mediated, by means other than getting reasons that support it. When he says that I must not suspect that I have willed myself to believe something, his reason for this does not depend on my having acquired the belief 'just like that' but only on my having done it in some way other than acquiring evidence for it. But Williams holds that there is no conceptual obstable to the mediated case, as in hypnotism or brainwashing: he objects to it morally, but thinks it coherent. Thus, his remarks imply the impossibility of something that he rightly thinks is possible.

IV WILLIAMS'S SECOND REASON

Here is Williams's second reason:

> Another reason stems from our considerations about perceptual belief: a very central idea with regard to empirical belief is that of coming to believe that P because it is so, that is, the relation between a man's perceptual environment, his perceptions, and the beliefs that result. Unless a concept satisfies the demands of that notion, namely that we can understand the idea that he comes to believe that P because it is so, and because his perceptual organs are working, it will not be the concept of empirical belief.... But a state that could be produced at will would not satisfy these demands, because there would be no regular connexion between the environment, the perceptions and what the man came out with, which is a necessary condition of a belief....

That is the whole of what Williams has to say on this score. (The omitted portions strengthen what is claimed, but add nothing to the argument for it.) From here he goes on to discuss the limited possibility, and reprehensibility, of *arranging* to acquire beliefs by the use of drugs, hypnotists, and so on.

The main premiss of this argument needs editing. The thesis cannot be that every empirical belief is held because it is true, for there are false beliefs. So Williams must be saying only that it is a conceptual truth that empirical beliefs *can* be acquired in ways that are explained by their truth. The word 'empirical', taken strictly, reduces this to the trivial point that immediately induced beliefs, just because they do not arise out of sensory experience in the standard truth-tending way, should not be accounted 'empirical'. I think, though, that Williams is using 'empirical', as philosophers sometimes do, alas, to mean 'contingent': his thesis concerns the concept of *beliefs about how the actual world is.*

With the premiss thus construed, I agree with it. It is a conceptual truth that (i) sensory encounters with the world are the standard, typical source for beliefs about how the world is, and (ii) in such an encounter the belief that P can arise from sensory intake from a state of affairs in which P — which is to say that the person believes that P because it is so. I would defend (i) on the basis of the broad rightness of the functionalist analysis of the concept of belief, according to which the concepts of belief and desire have as their essential role the explanation of how a creature's behavioral outputs relate to its sensory inputs. No sensory inputs, no analysis, and so no work for the concept belief. That does not give much centrality to (ii), but I don't think it is central in Williams's argument either, though it is perhaps covertly involved in his mention of a 'regular connexion' between 'the environment' and resultant belief.

What bearing does (i) have on whether beliefs can be immediately induced? That depends on what version of (i) we adopt. The version I would be willing to defend says that contingent beliefs must usually, centrally, and basically be acquired as a systematic result of sensory input from the environment. That makes no trouble for the possibility of immediately inducing belief. It does imply that immediately induced beliefs would have to be atypical or non-standard, but it does not rule them out. A state acquired in this way could count as a belief because of its relations — as exhibited in behaviour — with the creature's desires and its other beliefs. This does not kick off a vicious regress: *most* of the beliefs count as such because of their relations not to other beliefs but to the sensory inputs that cause them. But *some* beliefs may be acquired in other ways.

Williams thinks otherwise. He relies on the stronger thesis that contingent beliefs must *always* be acquired in that way: a state that 'could be produced at will', he says, could not be a belief because 'there would be no regular connexion' between instances of it and the person's sensory intake; and this has to mean that there would be no strictly, exceptionlessly regular connexion etc. That presupposes a version of (i) which does indeed conflict with the possibility of immediately induced belief, but I see no reason to accept (i) in that version of it. (All this applies equally to (ii), *mutatis mutandis*.)

My final point about Williams's first reason also applies to his second. The latter relies on the claim that it is part of the conceptual essence of belief that beliefs are always rooted in perceptual evidence, and the fact that if beliefs could be acquired voluntarily those roots would sometimes be cut. But this threatens to prove too much: if it holds against beliefs acquired voluntarily 'just like that' then it holds against beliefs acquired in any way that doesn't involve gathering evidence, e.g. through hypnotism. If Williams is right in holding the latter to be possible, his second reason cannot explain why immediately induced belief is impossible.

V SHARPENING THE PROBLEM

Each of Williams's explanations of why one cannot immediately induce belief fails because it gets no help from 'immediately'. If either of them did its intended job, it would also explain why one cannot induce belief in a mediated way; but there is no such fact to be explained.

Furthermore, there could be simpler, quicker, more reliable means for causing beliefs in people without giving them evidence. I passionately want to spend the evening in a state of confidence that the weather will be fine tomorrow (I have my practical reasons), so I give myself the thought of tomorrow's weather being fine while snapping my fingers in a certain way, and sure enough I end up convinced that the weather will be fine tomorrow. We have no such fast, reliable techniques for producing belief without evidence, but they are not conceptually ruled out.

So we don't leave the realm of abstract possibility if we think of people as being able to give themselves beliefs by snapping their fingers or downing a pill. If there is a conceptual impossibility, as Williams and I think there is, it comes between being immediately induced *to believe that P* and being induced to *do something as a means to believing that P*; the latter, however swift and sure the means-end link is, involves a mediated inducement to believe that P. The thinness of the line between those two makes my problem look harder than ever; but when the line was drawn in the wrong place the problem looked easier but was really insoluble.

VI THE BASIC TRIANGLE

The sought-for explanation must have something to to with the basic triangle: what a creature (1) does depends on what it (2) thinks and (3) wants. A full grounding of the concepts of belief and desire require this triangle to be developed into a square: (2) beliefs and (3) desires mediate between (4) sensory inputs and (1) behavioural outputs. Item (4) briefly raised its head in section IV above, but from now on I can set it aside and stay with the simpler triangular story.

Basically and initially it is a story about beliefs of the form: *Doing A is a way to get G.* Beliefs on other topics must be understood as adjuncts to, or means to, or upshots of, beliefs about means to ends. Whatever the grounds for saying that a creature somehow contains the information that P, it doesn't believe that P unless that information is manifested, through its beliefs about means to ends, in its behaviour.

There is a useful formula that was presented by Braithwaite in a seminal paper that reintroduced into twentieth century philosophy the insights of Alexander Bain's that had been neglected and forgotten.[12] The wording, but not the idea, is mine:

> x believes that P ≡ x is disposed to behave in ways which would (or will) satisfy x's desires if it were (or is) the case that P.

This, satisfactorily, covers not only beliefs about means to ends (Swimming is a way to get food) but also ones that have some bearing on the former (The prey went into the water). The corresponding formula for desire is this:

> x desires that P ≡ x is disposed to behave in ways that x thinks will bring it about that P.

Each analysed concept occurs in the analysans for the other. That is inevitable, which is why the functionalist literature proclaims that we must explain belief and desire together, through how they collaborate to explain behaviour.

The above formulae only approximate to analyses of the concepts of belief and desire; neither is true as it stands. To bring them closer to the truth we must qualify and complicate them in many ways — for example by playing off strength of desire against degrees of belief (levels of subjective probability). None of those details matter here, however. All I need is that the above two formulae capture the main lines of how the concepts of action, desire, and belief relate to one another. If you reject that, I have nothing positive to offer you concerning belief and the will.

VII Inducement and Desire

The triangle prompts a question about whether desires are involuntary as beliefs are, and if so how that might be explained. I shall follow this up for a section, in the spirit of Hume's advice that when a search is proving difficult one should 'beat about the neighbouring fields', I want the two searches to be as closely isomorphic as possible. Just as we have been examining the inducement of *any* change in the subjective probability one assigns to a proposition, so now I want to explore the inducement of any change in how strongly something is desired. Belief and desire are both matters of degree, and my topic is the relationship of the will to changes in the degree to which something is believed or wanted.

---

[12] R. B. Braithwaite, 'The Nature of Believing', first published in 1932 and reprinted in A. P. Griffiths (ed.), *Knowledge and Belief* (Oxford: Oxford University Press, 1967). Alexander Bain, *The Emotions and the Will*, first published in 1859 and re-issued by University Publications of America: Washington, D.C., 1977. See especially pp. 568–573 in the reprint.

There is no problem about the mediated inducement of desires. In explaining why, I have to distinguish 'end desires' from 'means desires': I have a means desire for something if I want it only as a means to something else that I want; I have an end desire for something if I want it for itself and not because of what I think it might lead to. I shall take the two sorts of desire in turn.

One's end desires can change — witnessing an execution turns a man against capital punishment, a life-threatening illness shifts his relative valuing of time and of money — and someone might undertake such a process in order to get an end desire that he doesn't now have.[13] Lucrezia, for example, doesn't care about the welfare of others, but she comes to think that people who do care are more serene than she is, have richer friendships and like themselves better. She therefore sets out to become like them — actually to have the welfare of others at heart, really to want it as an end desire, nor merely to act as though she wanted it. If she achieves this, her individual good deeds will be motivated by altruistic concerns — her wish for the good of others will be an end desire — and yet she will have come to this point because of an inducement, a belief about means to ends. There is a certain piquancy about this situation, but no paradox. It is simply a case where an inducement works in a mediated way on someone's end desires, closely analogous to the mediated ways of getting someone to acquire a belief.[14] Mediated inducements to acquire means desires are not only possible but in many cases easy. Suppose you were crazy enough to offer to reward me for *coming to want to put my hand into this bucket of mud as means to something*, I could easily arrange to acquire this desire by rigging the situation so that putting my hand into the bucket *is* a means to something that I want.

So we come to the question of whether one could immediately induce desires, which I take to be the question of whether desire is voluntary. It seems intuitively right to say that it is not, and that there is something conceptually wrong with the idea of coming to desire that P, 'just like that', because one sees advantage in desiring that P. Of course I can immediately get you to want to raise your arm by convincingly offering you a reward for *raising it* but not by offering you a reward for *wanting to raise it*. Or so I say, expecting agreement, on the strength of my intuitive sense of how things stand. But here as with belief it would be good to have an argument for this conclusion, preferably in the form of an explanation of why the conclusion is true.

---

[13] 'The will can contribute indirectly, as though from a distance, to make something pleasing to us.' G. W. Leibniz, *New Essays on Human Understanding*, trans. and ed. P. Remnant and J. Bennett (Cambridge: Cambridge University Press, 1981), pp. 182f.

[14] Locke, in the *Essay*, gives a structurally similar example: knowing that smoking is good for their health, people doggedly train themselves to like tobacco! Any change in likes and tastes is a change in one's end desires.

Here is an unsound argument for saying that an end desire could not be immediately induced:

> (i) If someone is induced to ϕ, then for some R he ϕ's so as to make R obtain; that's what inducement is. It follows that if someone is induced to desire that P, then *for some R he desires that P so as to make R obtain.* (ii) If someone's desire that P is an end desire, then *there is no R such that he desires that P so as to make R obtain*; that's what an end desire is. The two are formally inconsistent. What better proof could there be of the absolutely impossibility of inducing an end desire?

If this argument worked, it would work even against the mediated inducement of end desires; but the latter is possible, so the argument must be defective. Its fault is that it trades on the ambiguity in 'He desires that P so as to make R obtain'. It could entail that he thinks (a) that his wanting P will lead to R, or (b) that P will lead to R. These are independent propositions; neither entails the other; they bring out the radical ambiguity in the sequence in question. In the above argument, (i) is true only in the sense of (a), whereas (ii) is true only in the sense of (b). So the argument fails.

The immediate inducing of means desires can be dismissed as a separate topic for investigation. Suppose that on the stroke of noon I come to want to put my hand into a bucket of muddy water so as to retrieve money from the bottom of it, this being a means desire that I didn't have at 11.59:59. This implies that at noon (i) I think that putting my hand into the bucket is a way to get money from it, and (ii) I want to get the money. To acquire that means desire, therefore, I must have acquired either (i) the causal belief or (ii) the end desire for the money. So the means desire could not be immediately induced unless beliefs or end desires can be immediately induced; so the immediate inducing of means desires has nothing new to offer us and can be dealt with automatically through the discussion of our other two topics.

First objection: 'Your new desire for the money might itself be a means desire.' True, but that doesn't block the overall line of thought if it is otherwise all right. Here is the revised version that shows how the objection is met: My new means desire must come either from (i) a new causal belief relating hand and bucket to money or from (ii) a new end desire for the money or (a new means desire for the money, arising either from (i′) a new causal belief about money in relation to something else X or from (ii′) a new end desire for X or (a new means desire for X, arising either from (i″) a new causal belief about X in relation to something else Y or from (ii″) a new end desire for Y or (a new means desire for Y....))). This has to terminate in either a new belief or a new end desire, so my argument finds its way through this labyrinth to the conclusion that I have drawn.

Second objection: 'It could be that all morning you wanted the money and thought you could get it by delving into the bucket, but not until noon did that belief and desire generate a downright *desire* to put your hand in the water. If that were so, the acquisition of the means desire would not involve acquiring either a belief or an end desire.' Yes it would. The case as presented is not intelligible unless before noon the positive case for reaching into the bucket is outweighed by other considerations (e.g. I have another project that demands clean hands, and I have no facilities for cleaning my hand after putting it into the bucket), and if at noon there was no change in my belief about hand-to-money or my desire for money, then there must have been a change in how I viewed the rival project; and the latter change must have involved a change of belief or a change of desire (or a change in my relation to some third project, with this arising from...). So the second objection fails. We really can set aside, as not a separate topic, the immediate inducement of means desires.

Coming back then to end desires: Could someone be immediately induced to want something?' I have criticized one bad argument for answering No, and have offered no other. Still, I stand by my intuitively based conviction that that is the right answer, as it is to the corresponding question about the inducement of beliefs. It is reasonable to suspect, indeed, that the conceptual facts that generate one answer have a considerable overlap with those that generate the other. Enormously different as belief and desire are from one another,[15] there are abstract structural differences between them — and symmetries in how they relate to one another and to behaviour — that suggest that if they are both necessarily involuntary it is for similar reasons in the two cases.

VIII BELIEFS AND DESIRES AS DISPOSITONS

The basic functionalist triangle confirms something that has been accepted by many philosophers, not all of them functionalists. H. Price, for example, wrote: 'Believing a proposition is ... a disposition and not an occurrence or "mental act", though the disposition is not necessarily a very long-lived one....'[16] That seems right. Beliefs are not actions or events or processes; rather, they are sets of mind that are theoretically describable in statements about what *would* be the case about the believer *if* such and such came to be the case. A belief is a function from desires to actions. (Strictly: a total belief state is a function from desires to actions. A single belief is a function from desires and other beliefs to actions.)

---

[15] They are more different than is commonly realized, I think. See Jonathan Bennett, 'Analysis Without Noise' (section 6), in Radu Bogdan (ed.), *Mind and Common Sense* (New York: Cambridge University Press, 1990).

[16] H. Price, 'Belief and Will', *Proceedings of the Aristotelian Society*, suppl. vol. 28 (1954) 1-26, at p. 15.

Lately this view about beliefs has run into trouble on the grounds that beliefs cause behaviour whereas dispositions cannot be causes. I deny that beliefs cause anything, and suggest that the contrary opinion arises from a misunderstanding of the undeniable fact that attributions of beliefs help to *explain* behaviour. Are the explanations causal? That question is not determinate enough, just as it stands, for me to have an answer to it, but really it doesn't matter. Nobody, I presume, would deny that facts about a thing's disposition can enter into causal explanations of its behaviour. I realize that I am being rather dogmatic about this matter; it's a big topic, and I defend my position elsewhere.[17]

So I contend that to believe that P, for any P, is to have a certain complex behavioural disposition — more specifically, a function from desires to behaviour. And desires are dispositions too: a desire is a function from beliefs to actions. Many desires are accompanied by categorical processes, some of them all too painfully present to consciousness, but they are not of the essence of desire.

Is that a clue to the solution to my problem? A structural feature that is shared by beliefs and desires promises well as the source of the involuntariness of each, and the promise is confirmed by the fact that we have no trouble believing in the voluntariness of all the kinds of intellectual act that can be associated with acceptance (see section II). When we try to make good on the promise, however, things are not easy.

Here is an argument starting from the premiss that dispositions supervene on categorical (=nondispositional) states. This implies that one can create a disposition only by altering a subvenient categorical state — making something fragile by altering its crystalline structure, making it edible by cooling it, and so on. To give someone a belief, then, one must alter some categorical state of him upon which his having of that belief supervenes. So the inducing of a belief cannot be immediate, and must go through the inducing of an action — a change of categorical state — upon which the change of belief supervenes. End of argument.

The argument is open to the objection that it neglects the distinction between motivational and merely ontological mediation which I introduced in section I. I illustrated it through events that are causally required for some action to be done, whereas a change of categorical state doesn't cause the change of disposition that supervenes on it. Still, the relation between them is ontological – the world gets from the inducement to the change of belief through the subvenient categorical change – and it is a further question whether the person must aim to produce the subvenient change as a means to changing his beliefs. It has not been shown that he must. That is, it has not been shown that the categorical basis for a disposition yields not merely ontological but also motivational mediation. So the fact that inducements can change beliefs only in a motivationally mediated manner still waits to be explained. End of objection.

---

[17] Jonathan Bennett, 'Folk Psychological Explanations', in John D. Greenwood (ed.), *The Future of Folk Psychology: Intentionality and Cognitive Science* (forthcoming 1990).

The objection itself overlooks a certain distinction. Let us compare

> (i) the relation between an arm's being raised and the neural events that this involves

with

> (ii) the relation between a change of disposition and the categorical change on which it supervenes.

I have mentioned one similarity, namely that both are ontological, and one unlikeness, namely that (i) is causal whereas (ii) isn't. Another difference is this: It is not absolutely necessary that there be any neural or other preliminaries to an arm's going up, whereas every disposition absolutely must have some categorical basis, though it is a contingent truth that a given disposition has the specific basis that it does have. The first half of that is familiar as implied by broadly Humean considerations that have been widely accepted. The second half may need more comment.

I contend that although it could happen that two categorically indiscernible things were dropped in water and one dissolved while the other didn't, it couldn't be that the former was soluble while the other wasn't. This is because the statement that a thing *x* is soluble says something of the form

> If P were the case, *x* would dissolve,

which is true only if at all the closest P-worlds *x* does dissolve. Just what closeness is, for purposes of the theory of these conditionals, is not yet quite settled; but there can be no doubt that closeness somehow depends upon categorical similarity. We have here a richly promising line of analysis according to which dispositional differences must rest upon categorical ones, and I confidently accept that conclusion.

It follows that there could not possibly be such an event as *just* acquiring a belief. There are worlds at which, when someone raises her arm, her arm goes up without any neurological or other antecedents, but there are no worlds at which someone acquires a belief without undergoing a subvenient change of state. So the statement that when Henry acquired a belief that is all that happened is impossible, very much as it is impossible that two unalike things should become alike without either of them changing categorically.

I did not say 'change intrinsically', for there is the theoretical possibility that what someone believes might change because of changes not in his intrinsic nature but rather in how he is situated in the world.[18] It is hard to imagine someone setting out to believe that P by so altering his situation that one of his present beliefs comes to qualify as the belief that P, but if such a thing did happen, under inducement, this inducement would obviously be working in a tremendously mediated way. So from now on I shall ignore that possibility and attend only to changes of disposition that supervene on changes of intrinsic state.

We have absolute necessity here, but not equivalence. If x and y became alike because x lost its radioactivity, the proposition that they became alike is not equivalent to the proposition that x lost its ratioactivity, for the obvious reason that although the becoming-alike must supervene on *some* intrinsic change it need not have involved that one in particular, Similarly, attributing a disposition to a thing is not the same as attributing to it the disposition's categorical basis, because although the disposition had to have some categorical basis it didn't have to have that one.

Some may want to say that a particular disposition *token* is identical with its categorical basis. In my opinion most uses of 'token' in phrases like 'state token' and 'belief token' are vehicles for nonsense and covers for unthorough thinking. But I need not go into all that. My topic is what one can be induced to do, and thus what one can aim or try to intend or plan to do, and what one aims etc. to do is never to bring about a particular event but only to bring it about that there occurs an event of a certain kind.

So I stand by my claim that when someone is induced to acquire a belief, there must be something else that she is also induced to do. It is absolutely impossible for the belief to be acquired unless the person has a change of categorical state, but there is no such change C such that for her to be induced to believe that P is for her to undergo C. These are distinct states of affairs, yet one of them absolutely cannot obtain except as a consequence of the other or of some substitute for it.

IX THE FALSE ANSWER

That, I thought for a while, is why belief cannot be immediately induced. The line of thought went as follows.

In a strictly analogous way, I cannot immediately induce you to make two dissimilar things alike. It is absolutely, conceptually impossible for you *just* to make them alike; you have to make them alike *by*—through the mediation of—making an intrinsic change in one or both of them. You may even be able to do this without knowing in what ways they are unalike. In such a case, you will not aim directly at the basis for their becoming alike—that is, the intrinsic changes that will make them alike—but rather at something that will cause those changes. For example, you might set about making two documents alike by implementing a software program that compares them and edits one of them to match the other.

---

[18] In an earlier draft I did say 'change intrinsically'. John Hawthorne warned me.

Similarly with inducing beliefs. There is no reason in principle why some virtuoso of bio-feedback should not be able to focus his attention directly on the neural basis for acquiring the belief that P; but none of us can do this, for our own or other people's beliefs, and probably none of us ever will. What we have to do is to cause ourselves and others to have beliefs by aiming at something that will cause the belief by causing its neural basis: we subject the person to testimony or other evidence, or to brainwashing, hypnotism, drugs or the like, thus getting ourselves causally upstream from the neural basis and thus from the belief itself. In my account of the supervenience of beliefs upon intrinsic states, there is nothing to imply that to produce a belief in someone (myself or another) I must know anything at all about its neural basis. All that is implied is that I must know of something that is ontologically prior to the acquisition of the belief and must bend my efforts to the production of that. Whether it is prior by *being* or by *causing* the neural basis for the acquisition does not matter. What does matter is that it must be one or other of those. I cannot bend my efforts directly onto the acquisition of the belief.

This explanation of the involuntariness of belief has several prima facie virtues. (1) It is specifically about the *immediate* inducing of beliefs, and does not threaten to run wild and imply that beliefs cannot be acquired in any way except through the collection of evidence. (2) It applies to 'the acquisition of beliefs' when this is understood as meaning 'changes in subjective probabilities'. It does not apply to any intellectual acts whatsoever, e.g. to acts of acceptance or inner affirmation, if such there be. (3) In section V I contended that there could be extremely fast, simple, reliable ways for mediately inducing beliefs, and commented on how thin the line is between that sort of mediation and downright immediacy, remarking that that made my problem look harder than ever. But that harder-seeming problem is just what my explanation solves. The explanation says nothing about the length, difficulty or uncertainty of any means for producing beliefs; its whole point concerns the difference between changes of disposition (which do need a basis) and categorical changes (which don't). The line may be thin, but it is clear, sharp and absolute. (4) The explanation works equally well for the involuntariness of end desires. Indeed it can be generalized to argue that one could not be immediately induced to make *any* counterfactual conditional true, a result which I find intuitively acceptable.

X THREE OBJECTIONS TO IT

For all its seeming virtues, the explanation is a failure, as I now realize. Five of my friends have independently produced what amounts to the same objection.[19] I now state it in my own terms.

Whenever someone is induced to make it the case than an E event occurs, the person's thought must be in some degree abstract. I may aim to raise my arm in some fairly specific manner, but I cannot aim to raise it ϕly where ϕ is a complete specification of an arm-raising; the eventual, actual raising of my arm must have details that were not represented in my intention. So, with any voluntary or immediately inducible kind of action there is the same supervenience pattern that I have been emphasizing. The arm's going up supervenes, with absolute necessity, on the arm's going up in some specific fashion, and yet 'The arm goes up' is not equivalent to 'The arm goes up ϕly' for any completely specifying ϕ, because the arm could go up in any one of countless different specific fashions. So I have committed myself to the conclusion that there cannot possibly be such an activity as *just* raising one's arm, and the further conclusion that one cannot be immediately induced to raise one's arm. This is offered as a *reductio ad absurdum* of my explanation, and as such it resoundingly succeeds.

Perhaps my explanation was too parsimonious in its conceptual resources. Perhaps there is a sound explanation along those general lines—one that retains the prima facie virtues that I noted in section IX but relies on some further fact about (a) how a disposition relates to its categorical basis. It would have to be something that distinguished (a) not only from (b) the relation of an arm-raising to its neural cause but also from (c) the relation of an arm's being raised to its being raised in some perfectly specific manner. There may be something to work on here: In the case of (b) and of (c) the subvenient and supervenient items are of the same metaphysical kind, whereas in (a) they are not. But I have no way of firming and clarifying that, nor any suggestions for how it might help to answer my question.

Furthermore, any attempt along those lines would have to steer clear of a further objection that has been brought against my proposed explanation.[20] The route from an inducement to a change of disposition must be ontologically mediated; that much is agreed. I have argued that the 'must' expresses absolute necessity, and my explanation infers that therefore the mediation is not just ontological but motivational. It seemed plausible at the time, but now that it has been challenged I can find no reason to believe it. Motivational immediacy has to do with the agent's thoughts, whereas the absolute necessity of a categorical basis for a disposition resides in a conceptual connection which need not constrain the agent's thoughts because he may be unaware of it or outright disbelieve it. The premiss that *everybody must realize that* dispositions must have a categorical basis would offer some hope of securing my conclusion, if only the premiss weren't patently false. This second objection seems to threaten not only the explanation that it was aimed at but also any variant of it which stays at the level of conceptual and metaphysical connections.

---

[19] John Hawthorne, Mark Lance, Mark Brown, Frances Howard, and Alastair Norcross.

[20] By John Hawthorne, Frances Howard and Mark Lance.

If we leave that level, we presumably walk out on the project of explaining my explanandum by appealing somehow to some relation between a disposition and its categorical basis. A reason has been offered to me for thinking that if we can't get the explanation out of that we can't get it at all.[21] If something has to mediate between an inducement and the acquisition of a belief, and if the mediating item is not the belief's categorical basis, then it must be or involve a mediating event. But it cannot be absolutely necessary that *any* event mediates between the receiving of the inducement and the acquisition of the belief. It is always theoretically on the cards that receiving the inducement should lead in an ontologically (and therefore motivationally) immediate way to the acquisition of the belief. Let S_1 be the state of wanting to believe that P because one has just been offered a reward for believing it, and let S_2 be the state of believing that P. It may be biologically impossible, but it cannot be absolutely, conceptually impossible that someone should be wired up in such a way that S_1 *immediately* causes S_2. That is ontological immediacy, which entails motivational immediacy (or if it doesn't, I don't see why).

That threat to my explanandum can be countered. In explaining how, I am not mainly trying to protect my intuitive hunch that belief is essentially involuntary: the mere defence of intuitions is a trivial activity, and anyway this hunch of mine is not particularly strong. The importance of fending off the threat is that there is something to be learned from doing so.

I agree that the following is not conceptually ruled out:

A. Wanting to believe something could immediately cause one to believe it.

But I question whether that implies that beliefs could be acquired voluntarily. The view that it does follow presumably comes from the underlying view that

B. Doing something voluntarily is being immediately caused to do it by one's desire to do it, or something close to that. Because I reject B, I can find no plausible way of getting from A to the voluntariness of belief. B has been popular through the centuries, but it seems pretty clearly to be mistaken. If it were right, the following wish would be unintelligible:

---

[21] This objection came from Mark Lance and Frances Howard.

> I wish that my arm would rise right now without my raising it, going up simply as an immediate consequence of my wanting it to go up.

This seems to me perfectly intelligible, and I offer that as one way of seeing that whatever we mean by doing something voluntarily it is more than, or different from, it happening as an immediate consequence of wanting it to happen.

Still, the threat is valuable. It warns us of further complexities in our concept of voluntary conduct. The question of whether or why belief is essentially involuntary may be unanswerable until those further complexities are understood.[22]

Syracuse University,
Syracuse, NY 13244-1170, U.S.A.

---

[22] My debt to friends who have helped me with this work goes well beyond what is indicated in previous footnotes. All of them gave me other help also, enabling me to improve several earlier sections of the paper. I am truly grateful for their assistance.

PROGRAM EXPLANATION: A GENERAL PERSPECTIVE

By FRANK JACKSON *and* PHILIP PETTIT

SOME plausible assumptions generate a serious problem about the role of properties in causal explanations. This paper sets out those assumptions, identifies the problem they generate and makes a case for a particular solution.

The paper is in three sections. The first section introduces three plausible assumptions about the role of properties in causal explanations. We will not be questioning these assumptions here. The second section shows how these assumptions generate a problem when combined with a fourth assumption that looks equally difficult to resist. The third section offers a solution to the problem by showing how, nevertheless, we can and should resist this fourth assumption.

The fourth assumption is that the only way in which a property can be causally relevant to an effect is by being causally efficacious in its production. In order to resist that assumption, the paper draws on an account of causal relevance under which a property can be causally relevant without being causally efficacious. According to that account the realization of a property may program for the occurrence of an effect without actually contributing to its production. The paper, as the title suggests, provides another perspective on that programming account.[1]

I THE THREE BACKGROUND ASSUMPTIONS

The three following assumptions provide the backdrop for our discussion.

1. A causal explanation of something must direct us to a causally relevant property as opposed to a causally irrelevant property of the factor it identifies as explanatory: a property relevant to the causal production of the effect explained.
2. One way in which properties are causally relevant is by being causally efficacious. A causally efficacious property with regard to an effect is a property in virtue of whose instantiation, at least in part, the effect occurs; the instance of the property helps to produce the effect and does so because it is an instance of that property.
3. A property F is not causally efficacious in the production of an effect *e* if these three conditions are fulfilled together.
    (i) there is a distinct property G such that F is efficacious in the production of *e* only if G is efficacious in its production;
    (ii) the F-instance does not help to produce the G-instance in the sense in which the G-instance, if G is efficacious, helps to produce *e*; they are not sequential causal factors;
    (iii) the F-instance does not combine with the G-instance, directly or via further effects, to help in the same sense to produce *e* (nor of course, *vice versa*): they are not coordinate causal factors.

The first assumption hardly needs further paraphrase, though we should note that it is meant to hold good regardless of how causes are identified and individuated in particular cases. The second assumption is equally straightforward but it is worth remarking that the notion of efficacy it introduces is not tied to the view that causal efficacy is an irreducible feature of the world;

---

[1] This account is developed in Frank Jackson and Philip Pettit, 'Functionalism and Broad Content', *Mind* 97 (1988) 381–400, and 'Structural Explanation and Social Theory', in David Charles and Kathleen Lennon, eds, *Reductionism and Antireductionism* (Oxford: Oxford University Press, forthcoming).

it is compatible with a more or less debunking analysis of efficacy, say in terms of causal laws. One thing to note, however, is that no matter how the notion of causal efficacy is understood, it is distinct from the notion of instrumental effectiveness. A property will count as instrumentally effective *vis-à-vis* a particular effect, if it would have been a good tactic for producing the effect to realize that property. But such effectiveness does not entail efficacy: it does not mean that the effect occurred in virtue of the instantiation of the property.

The third assumption is less intuitively obvious than the other two and requires additional commentary. First, some elucidation. One way that F might be efficacious only if G is so, is by the F-instance and the G-instance being respectively more remote and more proximal causes of *e*; if they were sequential factors of this kind, both properties might be causally efficacious consistently with the truth of 3(i). 3(ii) is designed to eliminate this case. Similarly, 3(iii) is designed to make it clear that 3(i) is not true, just because the instances of F and G are each necessary parts of a causally productive complex of factors; if they were coordinate factors of that kind, the two properties might be causally efficacious consistently with 3(i). Notice that for all that 3 says, the instances of the distinct properties F and G may be identical. If G is efficacious in such a case, being a property in virtue of whose instantiation *e* occurs, still that will not make F efficacious: the instance of F will help to produce *e* but not because it is an instance of F; instead it will do so because it is an instance of G.[2]

What sort of situation would make 3(i), 3(ii) and 3(iii) together true? Well, here are some examples. In all of them the F-factor can be thought of as higher order and the G as lower order.

(A) A fragile glass is struck and breaks. Why did it break? First answer: because of its fragility. Second answer: because of the particular molecular structure of the glass. The property of fragility was efficacious in producing the breaking only if the molecular structural property was efficacious: hence 3(i). But the fragility did not help to produce the molecular structure in the way in which the structure, if it was efficacious, helped to produce the breaking. There was no time-lag between the exercise of the efficacy, if it was efficacious, by the disposition and the exercise of the efficacy, if it was efficacious, by the structure. Hence 3(ii). Nor did the fragility combine with the structure, in the manner of a coordinate factor, to help in the same sense to produce *e*. Full information about the structure, the trigger and the relevant laws would enable one to predict *e*; fragility would not need to be taken into account as a coordinate factor. Hence 3(iii).

---

[2] On related matters see Cynthia and Graham Macdonald, 'Mental Causes and Explanation of Action', *Philosophical Quarterly* 36 (1986) 145-58.

(B) I try and fail to fit a square peg in a round hole of diameter equal to the side of the square. Why did it not go through? First answer: because of the squareness of the peg. Second answer: because of the impenetrability of this overlapping part of the peg. The property of squareness was efficacious only if the overlap-cum-impenetrability-property was efficacious: hence 3(i) is true. But 3(ii) is also true, for the squareness did not help to produce the overlap-cum-impenetrability in the way in which it, if efficacious, helped to produce the blocking of the peg: there was no time-lag of the sort that such an influence would seem to require. And 3(iii) is also true, for the squareness did not combine with the overlap-cum-impenetrability to help in the same sense to produce the blocking; one could have predicted the blocking without reference to the squareness. As we might put it, the overlap-cum-impenetrability did not need any extra help from the squareness to produce the blocking.

(C) The water in a closed glass container reaches boiling temperature — the mean molecular motion is at such and such a level — and the container cracks. Why did it crack? First answer: because of the temperature of the water. Second answer, in simplified form: because of the momentum of such and such a molecule (group of molecules) in striking such and such a molecular bond in the container surface. (We are supposing that the case is one where the container breaks because of the internal pressure, not because of the temperature gradient between the water and the container.) The temperature-property was efficacious only if the momentum-property was efficacious: hence 3(i). But the temperature of the water — an aggregate statistic — did not help to produce the momentum of the molecule in the way in which it, if efficacious, helped to produce the cracking: hence 3(ii). And neither did the temperature combine with the momentum to help in the same sense to produce the cracking: one could have predicted the cracking just from full information about the molecule and the relevant laws. Hence 3(iii).

Is assumption 3 plausible? Well, given that the F and G properties do not relate as sequential or coordinate causal factors, it is certainly plausible that F cannot be efficacious in the same sense as G. Moreover it is plausible that if both are efficacious, then F is efficacious only in a derivative sense. The relation between the instantiation of F and the occurrence of *e* is secondary to the relation between the instantiation of G and that occurrence; other things being equal, the obtaining of the latter relation ensures the obtaining of the former, and not *vice versa*. But is it reasonable to go beyond these two plausible claims and endorse the claim in 3, that the F-property is not efficacious in any sense in producing *e*?

This is reasonable, and in two ways: strategically and theoretically. It is theoretically reasonable, because on any account of efficacy, it is Pickwickian to describe the F-property as efficacious, given that any efficacy it is alleged to have exercised would have been screened off by the influence of the G-property. No conception of efficacy, no matter how debunking, should allow that efficacy can be exercised across such a screen. Whatever the conception of efficacy in play, it must be admitted that if the instance of G helped to produce *e* because it was an instance of G, and if the instance of F was not a sequential or coordinate factor, then F cannot have been efficacious in the production of *e*: the instance of F cannot have helped to produce *e* or if it did — if it was identical, say,. with the instance of G — then it cannot have done so because it was an instance of F. And apart from those considerations, there is also this: that if higher-order properties are countenanced as efficacious, it seems we can invent efficacious properties at will. Not only was the fragility, the property of having a suitable molecular structure, efficacious in the breaking. So was the property of having such a property — if you like, meta-fragility; so indeed was the property of having that sort of property in turn — meta-meta-fragility; and so on into absurdity.

But it is strategically as well as theoretically reasonable to assert 3, rather than just admitting a derivative and a primitive sense of efficacy. Either being derivatively efficacious is a way for a property to be causally relevant or it is not. If it is not — if primitive efficacy is the only mode of causal relevance — then the problem to be raised in the next section remains, as will be there apparent. If derivative efficacy is a mode of relevance on the other hand, then while our problem goes away, it is replaced by a counterpart that requires the same sort of solution; it would require a solution like the proposal in Section III below. The counterpart problem is how to relate the two modes of relevance, how to make sense of the way different explanations of the same event can invoke properties that are efficacious at different levels: at the bottom level, primitively efficacious, and then efficacious at progressively more derivative levels.[3] For these reasons we shall not concern ourselves further with the notion of derivative efficacy and will take assumption 3 as given.

II THE PROBLEM

Our three background assumptions have a devastating impact once combined with the following proposition.

4. The only way for a property to be causally relevant to the production of a certain effect is by being causally efficacious in the process of production.

---

[3] This is a problem that arises, for example, for the theory presented in Peter Menzies, 'Against Causal Reductionism', *Mind* 97 (1988) 551–74.

Assumption 2 says that one way to be causally relevant is to be causally efficacious; this assumption adds that that is the only way. The combined impact of the four assumptions bears first on the possibility of causal explanation outside basic science and secondly on the possibility of causal explanation within.

The four assumptions entail that causal explanation in terms of causally relevant properties is only to be found within the realms of basic science: presumably, physics. For it is only in those realms that we seem to confront the sort of property which escapes being shown to be causally irrelevant by the combination of 1, 2, 3, and 4. This means that the four assumptions will drive us to dismiss the claims of the special sciences, and of course the claims of common sense, to be able to provide causal explanations in terms of causally relevant properties. For consider the sorts of properties invoked in those areas: the property of a group that it is cohesive; of a mental state that it is the belief that p; of a biological trait that it maximizes inclusive fitness. For each of these properties it is plausible that there is a property G lower down, so to speak, which is such that the higher up property F is efficacious only if G is, and yet the F-instance and the G-instance are neither sequential nor coordinate causal factors. But then by assumption 3, F is not causally efficacious, and by assumption 4 is not causally relevant. We shall have to regard all such properties as inefficacious, and so as irrelevant, and so as incapable of playing a role in causal explanation. The only properties with any claim to causal relevance and a proper place in causal explanation will be properties like mass and charge.

But even those who would happily turn their backs on common sense and the special sciences are going to be troubled by the combination of the four assumptions. Suppose that I explain the noise made by some mechanism by the property of the mechanism that some of its parts are loose. That property relates as F relates to G to the following more specific property: that this and that particular part are loose. It is the property, after all, of instantiating some such specific property, perhaps this, perhaps that, perhaps another one. Thus the explanation involving existential quantification — the reference to an indeterminate some — cannot be a proper explanation: it does not invoke an effacious property and so does not invoke a property that is relevant to the noise. The lesson holds quite generally, so that many of the explanations which physicists would endorse must look suspect. We cannot claim to explain the presence of vapour near the surface of boiling water by the fact that some of the water molecules have broken free. And we cannot claim to explain the radiation emitted by a piece of uranium by the fact that some of its atoms are decaying.

How might we hope to live with such results? They would mean that if in any area we gain access to a lower-order explanation of something, then we should jettison higher-order explanations in its favour. That in itself seems unattractive. We might want to give up the fragility explanation of why the glass breaks if we had access to the account in terms of molecular structure; that is unsurprising, since anyone who had access to the latter account would have all the significant information at his disposal which is offered by the fragility explanation. But would we want in parallel circumstances to give up the squareness explanation, the temperature explanation, the special-science explanations or explanations that make use of existential quantification? It seems highly doubtful, given that they offer us information which is not available just in virtue of having access to the lower-order counterparts. Someone who knows that the impenetrability of this part of the peg stopped it going through that hole does not necessarily know all that is known by someone who explains the blockage by the squareness of the peg. And so too for the other examples.

But we do not have access to lower-order explanations in many of these cases, so perhaps the assumptions are going to be palatable after all. Not so. The assumptions mean that we must view higher-order explanations, even when they are the only accounts available, as rough-and-ready stand-ins for explanations proper. The relation between such a stand-in and an explanation proper will be like the relation we might have thought existed between these two accounts of why the house is down: that it is because of the meteor that struck it yesterday; and that it is because of the event reported in today's newspaper. As we might have said that someone with knowledge only of the second account is certainly ignorant of the property of the explanatory factor in virtue of which it explains the destroyed house — the momentum of the meteor — so we will have to say that someone with access only to higher-order explanations of things will be constitutionally ignorant of the properties that explain the matters with which he is concerned.[4] For the properties he knows about are literally irrelevant.

All of this hardly makes for a satisfactory scenario and it hardly answers to our sense of how we are actually placed. Hence a problem. The problem is to find a way out, in particular to find a ground for rejecting the only one of our four assumptions which looks questionable: assumption 4. Can we offer an account of causal relevance which allows an inefficacious property to be relevant to the production of an effect?[5]

---

[4] Perhaps a more useful analogy is with two accounts of the property in virtue of which the meteor knocked the house: the false, folk account which holds the weight responsible and the proper account which refers us to the momentum of the meteor.

[5] For some recent discussions of this sort of problem see Simon Blackburn 'Losing Your Mind' in John Greenwood, ed., *Proceedings of the Greensboro Conference* (New York: Oxford University Press, forthcoming); Ned Block 'Can the Mind Change the World?' in George Boolos, ed. *Meaning and Method: Essays in Honor of Hilary Putnam* (Cambridge: Cambridge University Press, forthcoming); and Fred Dretske, *Explaining Behaviour* (Cambridge, Mass.: MIT Press, 1988).

III A SOLUTION

We can. In order to motivate the solution, consider a higher-order explanation involving existential quantification. Consider the explanation of why a piece of uranium emitted radiation over a certain period, which invokes the property of the uranium that some of its atoms were decaying: this, rather than the more specific property that such and such particular atoms were decaying. By our assumption 3, the property involving existential quantification cannot have been efficacious in producing the radiation. If it was efficacious, that is only because the more specific property was efficacious. And yet the instance of the abstract property did not relate to the instance of the more specific in the manner of a sequential factor or a coordinate one. So is there any other way in which the abstract property can have been causally relevant to the radiation, given that it was not causally efficacious?

Yes, there is, and the answer is more or less obvious. Although not efficacious itself, the abstract property was such that its realization ensured that there was an efficacious property in the offing: the property, we may presume, involving such and such particular atoms. The realization of the higher-order property did not produce the radiation in the manner of the lower-order. But it meant that there would be a suitably efficacious property available, perhaps that involving such and such particular atoms, perhaps one involving others. And so the property was causally relevant to the radiation, under a perfectly ordinary sense of relevance, though it was not efficacious. It did not do any work in producing the radiation — it was perfectly inert — but it had the relevance of ensuring that there would be some property there to exercise the efficacy required.

How are we to describe the relationship between such a property and an effect? The realization of the property ensures — it would have been enough to have made it suitably probable — that a crucial productive property is realized and, in the circumstances, that the event, under a certain description, occurs. The property-instance does not figure in the productive process leading to the event but it more or less ensures that a property-instance which is required for that process does figure. A useful metaphor for describing the role of the property is to say that its realization programs for the appearance of the productive property and, under a certain description, for the event produced. The analogy is with a computer program which ensures that certain things will happen — things satisfying certain descriptions — though all the work of producing those things goes on at a lower, mechanical level.

The solution proposed for the problem we have been confronting is that in each case the higher-order, inefficacious property is causally relevant to the event produced, because its realization programs for the realization of a lower-order efficacious property and, in the circumstances, for the occurrence of the event in question. The lower-order efficacious property may not be the lower-order property mentioned in each case but, if it is not, it will be one for which the realization of that property directly programs, one for which the realization of a property programmed for by that property directly programs, or whatever: it will be a property for which the original property programs indirectly, via the programming of intermediate properties.

The only way to bear out the proposed solution is to show that it plausibly applies in all the sorts of cases considered earlier. The fragility of the glass ensures, by the very meaning of what it is to be fragile, that the glass has a molecular structure — maybe this, maybe that — sufficient in the circumstances to produce the breaking. The squareness of the (impenetrable) peg ensures, as a matter of elementary geometry, that there will be an impenetrable part of the square end to obstruct its passage through the hole and again it may be this part or that which provides the obstruction. Finally, the temperature of the water more or less ensures — it makes it probable to a point approaching certainty — that a suitably situated molecule will have a momentum sufficient to break a molecular bond in the container and therefore to produce a cracking. What if the lower-order property mentioned is not itself an efficacious one? If it is not, then the story will go a level or more deeper until we find an efficacious property for which the original higher-order property programs indirectly, via the programming of the intermediate features.

It appears then that there are at least two distinct ways in which a property can be causally relevant: through being efficacious in the production of whatever is in question, or through programming for the presence of an efficacious property. The general problem raised in the last section can be solved by means of this observation, for the observation gives the lie to the troublesome assumption 4. It suggests with the sorts of cases used to illustrate the F–G relationship, for example, that the property playing the F-role has programmatic relevance to the occurrence of the event *e*.

Does the solution extend to other cases at which we have gestured? Well, it is clearly going to work with any explanations which are higher-order in virtue of using existential quantification. And equally, though we shall not explore the cases here, it seems to extend to explanations in common sense and the special sciences: for example, to explanations in sociology which invoke a property like group-cohesion, to explanations in psychology which invoke attitudinal contents as causally relevant properties, and to explanations in biology which appeal to a property such as that of maximizing inclusive fitness. In all such cases it is hard to see how the explanations can have the interest they clearly possess other than through being of the programming variety. The Yiddish *modus tollens* applies. If not this, what?[6]

Does the solution mean that in all of these cases we are entitled to provide the explanation we offer only if at some level there is an efficacious property at work for which the property programs? Yes, given there are efficacious properties, as our second assumption implies; but if there is an infinite progression of levels downward and therefore no effacious properties — by our third assumption — then the program story will have a different significance, bearing on relations between equally non-efficacious levels. Does the solution mean that we must be able to identify the efficacious property in question: that we must be able to provide the corresponding process explanation? No; it will do to have grounds for believing just that some such property must be in operation. In fact a little reflection suggests that perhaps most of the explanations we are ever likely to offer will be program explanations. Presumably we only reach potentially efficacious properties in physics. And presumably we will have access to these other than via existential quantification only in dealing with specific microphysical particles. But we will rarely be in a position to deal with specific particles and so most of the explanations we are ever likely to offer will be of the program variety.[7]

In conclusion, there is one point which it will be useful to emphasize. The notion of a programming property does not just explain how an inefficacious property can be relevant to the causation of an event. It also shows how a program explanation can have a significance that remains in the presence of an explanation invoking the corresponding efficacious property — the corresponding process explanation — and more generally in the presence of a lower-order explanation, whether it is of the program or process variety. A program explanation of an event *e* may provide information which the corresponding process explanation does not supply. Thus, it may be an explanation which the process explanation does not supersede.

The fragility case, as we have seen, is one where the claim fails: under natural background assumptions, we can say that someone who understands the molecular structure which accounts for the breaking of the glass understands all that is grasped by someone who offers the fragility account. The reason is that all it means to be fragile is to be such — say, to have such a molecular structure —

---

[6] For more detailed argument in the psychological and sociological cases respectively, see the papers mentioned in footnote 1.

[7] This is to take the program-process distinction as absolute. It can also be treated as a relativized distinction, with an arbitrary level of explanation being designated as involving causal process and then higher levels being cast as programming; in this sense, psychological properties would program for neurophysiological process, even if neurophysiological properties are not strictly efficacious. We take this approach in 'Structural Explanation and Social Theory'.

that breaking occurs under the relevant sort of knock. But the fragility case is exceptional. Someone who understands the lower-order explanations relevant in the other sorts of cases — even if they are process explanations — does not necessarily grasp the information available to someone who has access to the program explanations. If the molecular structure of a glass causes it to break in suitable circumstances, then the glass is fragile. But a part of the peg can stop it going through the hole without the peg's being square and the momentum of a water molecule can crack a container without the water's being at boiling temperature. Thus to know that the squareness and the temperature are explanatory, programming for the results in question, is to have information which is not available from the corresponding process explanations.

The point we are emphasizing can be put in other terms. According to David Lewis, to explain something is to provide information on its causal history.[8] Let us interpret the causal history as the process, involving such and such efficacious properties, that leads to the event or whatever in question. A program explanation provides a different sort of information from that which is supplied by the corresponding process account and therefore a sort of information which someone in possession of the process account may lack. The process story tells us about how the history actually went: say that such and such particular decaying atoms were responsible for the radiation. A program account tells us about how that history might have been. It gives modal information about the history, telling us for example that in any relevantly similar situation, as in the original situation itself, the fact that some atoms are decaying means that there will be a property realized — that involving the decay of such and such particular atoms — which is sufficient in the circumstances to produce radiation. In the actual world it was this, that and the other atom which decayed and led to the radiation but in possible worlds where their place is taken by other atoms, the radiation still occurs.[9]

Research School of Social Sciences,
Australian National University,
Canberra ACT 2601

---

[8] See the paper 'Causal Explanation' in his *Philosophical Papers*, Vol. 2 (Oxford: Basil Blackwell, 1988).

[9] This paper was written while Philip Pettit was a Visiting Fellow at Corpus Christi College, Oxford, with visitor's facilities at Nuffield College. He is grateful to both institutions for their support. We are also grateful for comments received on an earlier draft from Simon Blackburn, John Campbell, Jennifer Hornsby, Cindy and Graham Macdonald, Peter Menzies, David Miller, Michael Tooley, Tim Williamson and the members of a discussion group at Oriel College, Oxford.

WHY DID PSAMMENITUS NOT PITY HIS SON?

By AARON BEN-ZEEV

HERODOTUS ([5], 3.14) tells the story of the Egyptian king Psammenitus who did not weep at the sight of his son being led to his death, nor at the sight of his daughter being humiliated, but burst into tears when he saw his friend begging. Aristotle ([1], 2.8) says that this behaviour indicates that he felt pity toward his friend but not toward his son (and daughter); the son was too close to him to be pitied.[1] Why does the close relationship between Psammenitus and his son prevent Psammenitus from pitying his son? At first sight, it seems odd to say that a father does not pity his son — especially when he is being led to his death. Why should a basically positive attitude, such as pity, not arise in such relations? In order to answer this question, the nature of pity and its connection to the agent-object distance must first be discussed.

I

Pity may be described as a type of sympathetic sorrow over someone's substantial misfortune. Although this simplistic description captures some important features of pity, pity is a more complex attitude involving additional features. We may arrange the various features in light of the three major intentional components of emotions: cognitive, evaluational, and motivational components (see [2]). The features typical of pity are then the following.

(A) Cognitive features
   1. A belief that the object suffers from a substantial misfortune;
   2. A perceived impotence with respect to the object's misfortune;
   3. A belief that the object is somehow inferior to the agent;
   4. A belief that the object does not deserve such a substantial misfortune.

(B) Evaluative features
   1. Sympathetic sorrow over the object's suffering;
   2. A (tacit) fear of being in the object's situation;
   3. Satisfaction with the agent's position;
   4. Regret for some past activities.

---

[1] For some unknown reason Aristotle's version of the story has the father's name as Amasis who, according to Herodotus, was the father of Psammenitus. This substitution may be explained either as a slip of memory of Aristotle, or as a different version of the same story.

(C) Motivational features
  1. Acceptance of the present situation;
  2. Unwillingness to get personally involved;
  3. A desire for the object's relief.

One may organize the list differently or suggest a somewhat different list; I believe, however, that the above list reflects the major features of pity. Not all cases of pity include all these related features; however, prototypical cases do. The order within a certain group expresses the order of typicality: the first feature is more typical than the last one. I am not concerned here with the dictionary definition of the term 'pity', but rather with a certain psychological attitude toward someone's substantial misfortunes. I hope, however, that my analysis will capture the core meaning of the term. The purpose of dividing the above features into three groups is to indicate the dominant element in each group; this does not imply that other elements are not to be found in the group. Thus, the evaluational features include cognitive features and among the cognitive features there are also evaluative elements (especially in A4). I now examine each feature mentioned above and its connection to the agent-object distance.

The first cognitive feature suggests that pity does not emerge in every case of bad luck, but only when the agent believes that the object is suffering from a substantial misfortune. This feature is hardly connected to the agent-object distance. We are more aware of our intimates' suffering, but we may also be aware of the suffering of those who are personally unknown to us.

The second cognitive feature, namely the agent's perceived impotence with respect to the object's misfortune, is more related to the agent-object distance. Such a perceived impotence is more typical of distant relations. There are many creatures who suffer in the world, but one agent cannot do much to improve the situation of all of them; very often the most one can do (so at least many of us believe) is to help one's intimates. Thus, the suffering of homeless people is recognized by many people, but the majority think that they are unable to offer real help; the most they think they can do is to pay attention to the suffering of the homeless by pitying them. Accordingly, pity may be described as sympathy for the helpless by the powerless (or those who consider themselves to be powerless). Because of its non-interventionary nature, we speak about 'the luxury of pity'. Pity is improper if one has power to alleviate suffering. Doctors who can cure their patients do not pity them. Similarly, it is improper for the president who can help the homeless to pity them. Pity is also improper if my son is homeless and I have the resources to help him. The closer the object is to the agent, the more the agent is obliged to assist the object through hardship.

I wish to emphasize that pity is characterized by a *perceived* impotence rather than a real one. In many cases the agent can actually help but perceives himself as being unable to do so because of certain conceived (objective and subjective) constraints. Thus, although I can help a certain beggar by giving him most of my salary and devoting most of my time to him, I perceive this possibility as not practical in light of my obligations toward my family and my wish to keep a certain style of life. In such cases the perceived impotence is really a perceived lack of obligation which often stems from unwillingness to get personally involved (C2). There are, however, other cases of pity (for instance, in the case of a man dying of cancer) in which the perceived impotence expresses a real impotence. It is clear that the impotence in the first cases is closely connected to the agent-object distance.

Pity is morally valuable since it involves paying attention to the suffering — rather than the success — of others; it overcomes our natural tendency to look away from those who suffer and it may lead to some forms of help. However, mere acknowledgment is not enough in very close personal relations; real help is required. In cases like that of Psammenitus where the agent cannot offer real help even in close relations, the attitude is closer to pity than otherwise in close relations.

Turning to the third cognitive feature, pity often involves the agent's (frequently tacit) belief that the object — at least in his present situation — is in some way inferior to the agent. This feature (which is emphasized in Nietzsche's conception of pity) is absent from compassion; compassion involves transcending various types of inequality and assuming equality in common humanity (see [3]). In close relations compassion is more common than pity. It should be noted that although the difference between the prototypical cases of pity (in the sense used here) and compassion is quite significant, the difference between them is in degree rather than in kind. Since most of the features which are typical of pity can appear in various degrees, a change in the degree of certain features can turn the attitude from pity into compassion. (I further discuss A3 while discussing the features associated with it: B3 and C1–2.)

The fourth feature in the cognitive group, namely the agent's belief that the object does not deserve such a substantial misfortune, is less universal. Thus, I can feel pity toward a mass murderer although I believe that he deserves to be executed. However, it is often the case that, even if the agent believes that the object deserves some kind of punishment, by pitying him the agent expresses the belief that the misfortune is too severe. In a close agent-object relation, the agent would more easily believe that the object does not deserve such a misfortune. (Although the evaluative element in this feature is quite central, I do not include it in the evaluative group because I want to distinguish between this feature and the emotional attitudes included in the evaluative group.)

The second group of features is concerned with the isolated emotional evaluations underlying the complex attitude of pity. An emotional attitude does not usually consist of merely one simple emotion, but rather involves a complex pattern of emotions. Grief, for example, may involve anger, guilt, and shame; and love may include joy, jealousy, hope, admiration, etc. This group expresses the agent's evaluative stand toward the object; thus sorrow expresses a certain disapproval of the situation (or some aspects of it), and hence the agent's negative evaluation of it. Similarly, the unique nature of envy may be described as involving a type of negative evaluation of the good fortune of others. Each emotional attitude may be defined by its underlying evaluative attitude. For the sake of simplicity, I do not analyse the complex evaluations underlying each emotional title here, but merely mention the title while assuming the reader to be familiar with the basic evaluative stance in question.

The first feature in the evaluative group (like the first one in the cognitive group) is typical of all (or most) types of sorrow and does not presuppose a particular distance in the agent-object relation. I can feel sorrow for those who are very distant from me. (The intensity of the sorrow depends, of course, on the agent-object distance.) Pity then is a particular type of sorrow, and we like to know whether this type is sensitive to the agent-object distance. The other features of this group are indeed not typical of all types of sorrow; furthermore, their emergence and nature (and not merely intensity) are quite sensitive to the agent-object distance.

The second feature in the evaluative group involves the agent's fear that she, or some people who are close to her, may suffer such a misfortune. Typical pity involves the agent's imaginative reconstruction with the agent himself in the object's situation. This reconstruction is easier, the greater the psychological proximity. It is unlikely that we pity those who are very dissimilar to us since it is hard to imagine ourselves in their place. Despite the easiness of reconstruction, this feature is not dominant in very close relations where the central concern is what happens to the object rather than the future state of the agent.

The next feature in the evaluative group is a feeling of satisfaction which is based upon the favourable comparison between the agent's and the object's situations. Accordingly, pity sometimes involves the agent's thought (and maybe even the pleasure) of being lucky. If the agent thinks that his position also reflects credit on him, then the satisfaction will involve pride too. The agent's satisfaction, which is compatible with the belief in the inferiority of the object, is clearly absent in very close relations. This feature of satisfaction and the first two motivational features (C1, 2) seem to be the most important features of pity which are absent in very close relations.

In light of the presence of satisfaction (which is accompanied by the superiority feeling), pity may easily insult or humiliate the recipient. This is why most people do not like to be pitied. Accordingly, Kant and Nietzsche, who assume that pity is worthless from a moral viewpoint, argue that if one cannot overcome this emotion, one should not allow others to know that they are pitied (see [4], [5]). (Some people like to be pitied mainly because of the attention they get and which otherwise would not be given to them.) The feature of satisfaction is expressed in La Rochefoucauld's saying that 'There is in the misfortunes of our best friends something that is not displeasing to us'. For similar reasons grief may sometimes be pleasant — not on its own account, but owing to some ideas associated with it. The feature of satisfaction in pity is somewhat similar to the feature of delight found in the emotion of pleasure-in-others-suffering (*Schadenfreude*), but the two complex emotional attitudes are quite different. Whereas pity is basically *sorrow* over the object's *substantial* misfortune, pleasure-in-others-suffering is basically *delight* over the object's (usually) *minor* misfortune.

The last feature of the evaluative group — the agent's regret over some of her past behaviour — is again more evident in close relations where the agent could change the course of events which leads to the object suffering.

With the exception of the first feature (which is typical of all types of sorrow), the emotional components which are typical of pity are sensitive to changes in the agent-object distance. Pity involves the agent's viewing the object as another person who is quite separate from the agent. This is not typical of very close relations where the agent feels to be closely associated with the object and often has emotions similar to those of the object. When an unjustified misfortune falls on one who is very close to us we feel anger rather than pity. As Patricia Greenspan ([6], p. 66) rightly argues, the difference between anger and pity is not just that pity lacks the intense feeling dimension of anger, but that 'it also lacks anger's orientation towards change'. Pity, 'is compatible with a kind of acceptance of the target's situation that is not compatible with anger'. This type of acceptance underlies the motivational features of pity.

The motivational group is the most sensitive to the agent-object distance. Accepting the object's situation and being unwilling to get personally involved are typical only of distant relations. Thus, a father of a retarded child may be angry at his son's misfortune or feel ashamed of him, but his emotional attitude will not be that of pity. The father's attitude would not be that of having nothing to do with this inferior creature, but rather that which emphasizes the human features of the son, and that which involves helping the son as much as possible. The agent's acceptance of the object's situation and his unwillingness to get personally involved are often associated with the agent's perceived lack of obligation. Such a passive motivational stand may either stem from (a) the agent's belief that the object is somehow responsible for his inferior position or that his position is unalterably inferior, or (b) the agent's perceived lack of the required resources. Pity is more intense in the second case where the object's relief is possible in principle, but is not feasible under these particular circumstances. Generally speaking, a near success is more frustrating than a simple failure, and a misfortune appears less negative when it is perceived as necessary rather than something that could easily have been prevented (see [7], pp. 141–4). Compassion involves a greater commitment to personal help than that found in pity. In compassion, like in anger, the agent expresses dissatisfaction with the present situation. Yet unlike anger the agent conceives of it as unalterable in the short run; hence, the agent tries to help the object endure the painful situation rather than help the object change it. Even this kind of acceptance is absent in anger. Anger involves the desire, and often the intent, to change (and even destroy) this situation in the short run.

The last motivational feature, namely desiring the object's relief, is connected with whether the object deserves this misfortune. The typical belief is that the object does not deserve such a misfortune and hence the agent usually desires the object's relief. This desire is more evident in close relations.

To sum up, we have found in each group two major features which are usually absent in very close relations — A(2, 3), B(2, 3) and C(1, 2). If we take into account that A1 and B1 are typical of all (or most) types of sorrow and that our purpose here is to find out whether a particular type of sorrow — namely, pity — is typical in close relations, the absence of the above characteristics becomes more important. Actually, only the last feature in each group (which is also the least typical feature) can be counted as an indication that the type of sorrow which is typical of very close relations is pity. In light of the absence of the other more typical characteristics this is not a good indication at all. I do not want to argue that the agent-object distance is the only factor in classifying a particular sorrow as a pity, but I have suggested that in some situations this is a very important factor for such a classification.

II

We are now in a position to explain Psammenitus' behaviour. Aristotle ([1], 2.8) explains such a behaviour by saying that, 'we pity those who are like us in age, character, disposition, social standing, or birth; for in all these cases it appears more likely that the same misfortune may befall us also.' Pity then presupposes a certain proximity which enables us to relate ourselves to the emotional object and to realize his misfortune. However, the proximity should involve some distance; we do not pity, Aristotle suggests, those who are very close to us.

The above discussion of pity suggests why Psammenitus did not pity his son when he saw him being led to his death. Psammenitus may have had A(1, 2, 4), B(1, 4) and C(3), but not the other features. (Again, the important features of pity which distinguishes it from other types of sorrow are missing from this list.) Psammenitus believed that his son was suffering from an undeserved misfortune with respect to which Psammenitus was impotent; Psammenitus had sorrow over his son's suffering, and may have had regret for some of his past activities; he also desired his son's relief. But in no way did Psammenitus' attitude involve a type of acceptance and justification of the son's situation — as is often the case with pity. Psammenitus had the same attitude as his son, namely the attitude of anger and fear typical of unjustified threatening situations. The dominant attitude of Psammenitus was neither that of recognizing the inferior position of another person and wishing not to be in such a position, nor that of satisfaction with his own superior position. Rather the attitude involved realization that a very important part of himself was being destroyed; therefore Psammenitus was terrified and angry. The son's inferior position was experienced as Psammenitus' own inferior position. The dominant wish of Psammenitus was not to avoid such a situation in the future, but to change the present situation.

Psammenitus explained his behaviour by saying: 'my own griefs were too great to cry out about, but the sorrow of this friend is worth tears; he had much, and much happiness, and has lost all and become a beggar when he is upon the threshold of old age' ([8], 3.14).[2] Psammenitus' explanation refers to his children's misfortunes as his own; hence, he does not pity them. In the case of his friend, the agent-object distance was sufficient for pity to emerge. The pitier must feel close enough to the pitied person — but far enough away not to fall within the range of other emotions such as anger or fear. Since Psammenitus was very close to his son, crying while seeing the son might be interpreted as dishonourable behaviour of begging for his own sake; such an interpretation is unlikely in the friend's case because of the distance between them. (There may be other behavioural explanations to Psammenitus' cry in one case and its absence in another — thus, it may be argued that the friend's begging is less expected and more humiliating; but our concern here is not with the behavioural phenomena of crying but with the emotional attitude of pity.)

---

[2] Herodotus tells that Cambyses, the Persian king who ordered the execution of Psammenitus' son, was so impressed by Psammenitus' behaviour that he ordered that the boy's life should be saved. It was, however, too late: those who were sent after the boy did not find him alive.

Psammenitus' lack of pity for his son was not due to a lack of sorrow at seeing him being led to his death. Both pity and Psammenitus' attitude include more than just a general sorrow. But the additional features are different in each case. Psammenitus' attitude is not merely devoid of many features typical of pity, but it includes novel features as well. It is a different complex attitude.

Psammenitus' behaviour is unique. Most people would burst into tears in a similar situation. (Tears are neither a necessary nor an exclusive expression of pity.) Nevertheless, their emotional attitude would not be pity. Their emotional attitude would be a complex which includes sorrow, but often not pity. In the cases where the complex would include a component of pity too, this component would not be the central one.

We can describe the relation between the likelihood of the emergence of pity and the agent-object distance as having a bell-shaped curve. In very small and great distances these attitudes do not usually emerge; they are typical of a medium-size distance. Such a curve is typical also of other emotions. Thus, envy is similar to pity in this respect. We envy those who are like us or equal to us, but we usually do not envy those who are very close to us or very far from us.

III

After suggesting that very close relations may prevent pity, I would like to briefly discuss the phenomenon of self-pity which would appear to contradict this suggestion since the relation between the two variables in self-pity seems to be the closest one can get: they actually consist of only one entity. The problematic nature of self-pity is illustrated in Aristotle's view where there is no place for such an attitude. Aristotle's definition of pity contains two major features: (a) a feeling of pain at an apparent evil which befalls one who does not deserve it, and (b) expectation that this evil will befall upon ourselves (or some friends of ours). Self-pity contains the first, but not the second feature. Accordingly, Aristotle does not speak about self-pity (see [1], 2.8; and [9]).

In my view, prototypical pity contains more features; however, not all of them have to be present in each instance of pity. In self-pity the agent creates the agent-object distance (which is required for pity) by detaching herself from her present situation and viewing herself as though she were someone superior to the person she actually is now. The present position is viewed as the consequence of some arbitrary misfortune and does not reflect her 'proper' position. In light of this imaginative reconstruction, self-pity contains many typical features of pity. In contrast to the attitude typical of very close relations, self-pity involves the agent's perceived impotence, her acceptance of the situation and her unwillingness to get personally involved. Indeed, the agent's pronounced self-impotence is what gives this attitude its negative connotation. The agent talks all the time about her misfortune, but does nothing to change it. The only two features which seem to be absent from self-pity are the satisfaction with the agent's position and the fear of being in the object's situation. (But it may be argued that self-pity is not completely devoid of these features too. Self-pity may be viewed as a way of protecting the agent's positive self-image by not attributing the present inferior situation to her, and as involving the agent's fear of being permanently in this inferior position.) Thus, self-pity involves many typical features of pity, and is a type of pity – although not the prototypical one.

The possibility of self-pity and the unlikelihood of pity in very close relations indicate that the relevant agent-object distance is a constructed psychological distance rather than an actual physical separation. The psychological distance is very flexible and sensitive to contextual features; this is one reason for the complexity of pity.[3]

University of Haifa,
Haifa 31999, Israel

REFERENCES

[1] Aristotle, *Rhetoric.*
[2] Aaron Ben-Zeev, 'The Nature of Emotions', *Philosophical Studies* 52 (1987) 393–409.
[3] Lawrence Blum, 'Compassion', in *Explaining Emotions*, edited by A. O. Rorty (Berkeley: University of California, 1980), 507–17.
[4] Eamonn Callan, 'The Moral Status of Pity', *Canadian Journal of Philosophy* 18 (1988) 1–12.
[5] David Cartwright, 'Kant, Schopenhauer, and Nietzsche on the Morality of Pity', *Journal of the History of Ideas* 45 (1984) 83–98.
[6] Patricia Greenspan, *Emotions and Reasons* (New York: Routledge, 1988).
[7] Fritz Heider, *The Psychology of Interpersonal Relations* (New York: Wiley, 1958).
[8] Herodotus, *The History* (Chicago: The University of Chicago, 1987).
[9] Michael Stocker, 'Affectively and Self-Concern: The Assumed Psychology in Aristotle's Ethics', *Pacific Philosophical Quarterly* 64 (1983) 211–29.

---

[3] I am grateful to Stephen Toulmin, David Rosenthal, Lisa Frank, Irwin Goldstein, Peter Smith, Peter Stearns, Michael Strauss, Ruth Ben-Zeev and an anonymous reviewer for helpful comments. The paper was written while I was at the Center for the Philosophy of Science at the University of Pittsburgh, and the Center for Interdisciplinary Research at the University of Bielefeld. The support of these institutions is gratefully acknowledged.

ACTION, UNCERTAINTY, AND DIVINE IMPOTENCE

By TOMIS KAPITAN

IF deliberation aims at settling the mind upon a course of action, then a deliberator must be partly unsettled prior to decision. There is no point in deliberating whether to do what one is already committed to, regardless whether the commitment is practical or doxastic. One might reconsider a decision, or overlook it while deliberating anew, but each case reinstates non-commitment. Alternatively, one can anticipate that one will perform an act if one chooses or if one implements some broader action plan, or, convinced that one will perform the act, one might consider how to do it. This involves deliberating 'about' the act, but not whether to perform it. Deliberation seems wedded to antecedent uncertainty, making it impossible that a deliberator comprehends its future in full detail.

Does a similar suspension of commitment antedate all intentional action? Assuming that intentional action is the outcome of an agent's intending something or other, and that intending is a matter of 'settling' the mind (see [1] and [11]), then prior uncertainty is arguably essential to intentional action. Of course, not all anticipation is precluded; we inevitably act upon some survey of the future (and the past), but the scope of our vision is limited. Ignorance underlies not only our freedom of decision, but our very capacity to act, making it impossible to pair agency with omniscience. To establish these harsh conclusions is the object of what follows.

I INTENTION

If an action or doing is an agent's performance of, or refraining from, an action-type at a time, we distinguish those that are intentional from those that are accidental or instinctual. One exercises agency only through intentional action, and an action is intentional only if among its causal antecedents is an intending of the agent's — where an intending is a state whose content is an intention (see [3]). This does not imply that whatever is intentional is intended (see [1], [11]), nor that every intending is deliberative; it indicates simply that what one does intentionally is connected, at least causally if not also conceptually, to what one intends. Though not a sufficient condition for intentional action, it fails to be even necessary for accidental or instinctual doings.

Yet, talk about 'intending' or 'having an intention' is ambiguous. A distinction between having an intention (being inclined in a certain way) and consciously intending (voliting) is obvious enough. The former is not just a propensity to act; a baby does not intend to suck, despite its evident propensity to do so. At best, intendings are a subclass of propensities, distinguished by the fact that one must come to intend at some point in time. One cannot intend without having adopted or acquired an intending attitude, and this requires a change in behavioural proclivities if we are to demarcate between intentional and non-intentional activity.

We thereby distinguish having an intention from acquiring an intention and both, in turn, from rehearsing that intention, viz. consciously affirming an intention already held. Last night I decided to swim at nine o'clock this morning and, upon waking, I again thought, in an intending way, of swimming at nine. I did not make up my mind anew, but simply rehearsed an antecedently acquired decision.

II INTENTION AS LEAST EFFORT: EFFICACY

No one acts intentionally from a zero point of view, devoid of antecedent knowledge or beliefs. Every intention is acquired against a background of beliefs, plans, desires, and other intentions, with some idea about the probable role and contribution of the intended action. None of us intends to swim at the local YMCA without assuming that there is a local YMCA or that it has a pool. We do not expend the mental effort if we are cognizant of the utter futility of so doing.

Reference to an expenditure of effort is helpful. To intend is to ready oneself for the exertion of effort, and often, is itself the first phase of such exertion. If nature does nothing in vain, an analogous principle of 'least effort' affects intention-acquisition; we acquire an intention to A only if we assume that we are not wasting our effort, i.e., that our A-ing is open for us and, therefore, (i) that intending to A is something that needs to occur if we are to realize certain ends, (ii) that so intending will not be obviously fruitless, and (iii) that it is not otherwise inevitable.

In familiar terms, this requires that we take the course of action to be 'under our control', allowing that the appropriate belief be tacit and comparatively lower-level. Included is not only a belief in the presence of appropriate means but also a tacit presumption of efficacy; one intends to A only by assuming there is some probability of success, that is,

(1) If, at t, X acquires an intention to A at t′ then X believes that there is a chance that he would A (would refrain from A-ing) at t′ were he to intend to A (to refrain from A-ing) at t′.

Here, the projected time of intending must fall within the interval bounded by t and t′, and the weak probability qualifier may be deleted if X believes his effort is guaranteed success, or if the embedded conditionals harbour a tacit *ceteris paribus* concerning the improbability of unforeseen intervention.

But satisfying the consequent of (1) is not sufficient for taking A to be under one's control; the 'least effort' model also precludes viewing A-ing as *fait accompli*, i.e.,

(2) If, at t, X acquires an intention to A at t′ then X believes that he himself would not A at t′ unless he intended to A at t′.

To acquire an intention, in short, is to sense its need, to realize, at least tacitly, that what is intended is not otherwise inevitable. Even though I might deliberate about something I know I am going to do regardless of what I intend — for example, crash into the concrete barrier that suddenly looms before me — I deliberate only about how to do so, at full speed, say, or intentionally, for here is something that remains to be fixed.

III INTENTION AS LEAST EFFORT: CONTINGENCY

There is more to the presumption of openness and to the 'least effort' principle than (1) or (2). Nathan might think he would drink whiskey were he to try, would not do so unless he intended, yet also believe that he will *not* so choose. He has ruled out drinking whiskey, and having done so his mind is already settled and he no longer considers the action open, a fact which cannot be explained by the conditional consequences of (1) and (2) alone. Drinking whiskey is open only if Nathan takes it to be possible that he so intend and possible that he does not, that is, only if he believes that his decision about the matter, whatever it might be, is as yet contingent.

Mere logical or nomological contingency is too weak for interpreting this embedded modality; instead, a suitably concrete relative modality is at stake, where P is contingent relative to a set S just in case neither P nor not-P is a consequence of S. But how shall we specify the set S relevant to action? Three candidates are paramount:

(a) all propositions true at t (including those with reference to the past and future);
(b) all states of affairs (facts, conditions) obtaining prior to and including t; or,
(c) all that he (the agent) then (at t) believes.

Each of (a)–(c) has been advocated, though it is unimportant for present purposes which is preferred.[1] The doxastic modality of (c) has the advantage of not rendering a decision-making determinist inconsistent, and nothing beyond it need be assumed here. Still, (c) can be understood in at least two ways depending on whether the phrase 'he then believes' occurs outside or inside the scope of belief.[2] The first-person character of intention suggests an internal occurrence, i.e.,

---

[1] If we are speaking of action, and not just of choice and intention, then analogues of (a)–(c) are found throughout the literature; for instance, (a) is found in [6], (b) in [15] and [16], and variants of (c) in [5]. I have argued in [10] that (c) is all that is required for the sense of openness, and it should be noted that for an omniscient agent, (c) is sufficient for (a) and (b).

(3) If, at t, X acquires an intention to A at t′ then X believes that his intending to A at t′ is yet contingent relative to what he himself then believes.

The result equates intending with choice. Thus, by conjoining the consequents of (1), (2) and (3) we obtain a promising analysis of what it is for an agent to take a course of action as open. If choosing is nothing more than intending one course of action from a range of open alternatives, and a minimal range is the pair comprised of X's doing A and X's refraining from A, then a minimal choice is a selection from a minimal range. It follows from (1)–(3) that every instance of intention-acquisition is a minimal choice.

IV THE ANTECEDENT SENSE OF OPENNESS

While the consequents in each of (1)–(3) attribute belief, the times of these beliefs remain to be fixed. How are they to be identified in relation to the times of action and intending?

Undoubtedly, the time of believing is no later than the time of intention-acquisition. But they cannot be simultaneous, at least not to an agent who is minimally rational (locally consistent in the relevant beliefs) and self-reflective enough to realize he has intended when he in fact has. That is,

(4) If, at t, a minimally rational agent X both acquires an intention to A at t′ and believes that he has, then at some time t″ appropriately prior to t, X holds the belief attributed in the consequent of (3).

Proof. Assume the antecedent and suppose that the only time at which X held the beliefs in the consequent was simultaneous with t. Realizing that he has at t the intention to A at t′ then, relative to his own beliefs at t, his intending to A at t′ is no longer contingent, and, being rational, he could not at t take it to be so, i.e., X cannot believe that he will intend to A while at the same time view his intending to A as contingent relative to what he believes. So, the time at which he takes his intending to A to be contingent must be appropriately prior to time t — where a commentary on 'appropriately prior' awaits a more detailed study of the temporal parameters of practical thinking. It follows that X cannot come to intend except by antecedently presuming that his intending to A is contingent relative to what he then believes.

---

[2] See [4] for a nice discussion of the contrast between external and internal occurrences of terms within attitudinal scope. Also, where x is omniscient we have (u) (x believes [P]) iff x believes [(u)P], regardless of the restriction on u.

V THE PRESUMPTION OF UNCERTAINTY

While it is wrongly assumed that a deliberator cannot know in advance what he will do, decision does involve passing from a state of uncertainty into knowledge (see [7], [9], [10], [12]). Of course, it strains the imagination to speak of deliberating whether to A while occurrently believing that one will A, still more if we suppose the deliberator is occurrently intending to do A; 'whether' suggests a moment of uncertainty and indecision which precludes having one's mind made up modulo occurrent conviction. Perhaps a deliberator's antecedent indecision, ignorance, or lack of doxastic commitment can be restricted to occurrent states. But this would be of no help in understanding the state of uncertainty which, like wonder or bewilderment, harbours intentionality. Conditions that prohibit occurrent intendings or believings are expressed with negation having largest scope; ascribing no content at all, they fail to account for an agent's sense of uncertainty. It is the contentual aspect of the latter which demands scrutiny.

A deliberator who responds to a question about what he will do may very well say; 'I don't know, I've not yet made up my mind,' implying not indecision, only a presumption of such. For a minimally rational agent satisfying the consequent of (2), this is a consequence of a further presumption of uncertainty; if the agent believes that he himself does not yet know whether he will intend to A at t, then, by (2), he believes he does yet know that he will A — hence, a presumption of indecision. But such uncertainty is itself a result of the presumed contingency:

(5) If, at t, X believes that his intending to A at t′ is contingent relative to what he himself then believes, then, at t, X believes that he does not yet believe that he will intend to A at t′ and he does not yet believe that he will not intend to A at t′.

Proof. Suppose X satisfied the antecedent. Then, if he takes his intending to A at t′ not to be contingent with respect to a set S he will not regard S as a set of his own beliefs. Given the rationality proviso, he does not suppose his intending to A to be contingent with respect to any set among whose members is the proposition that he intends to A at t′. So, he does not view the latter as a set of his own beliefs, that is, he presumes that he himself does not believe he will intend to A at t′ and, similarly, that he himself does not believe he will not intend to A at t′. Therefore, he satisfies the consequent.

VI A THEOREM CONCERNING OMNISCIENCE

Satisfying the consequent of (5) does not imply that X does not know, or have a belief about, what he will intend to do, for the attributed belief might be false. An exception to this qualification occurs when the agent is omniscient.

Let us understand by 'God' an omniscient being. Then the following equivalences hold:

(6) For any proposition p and time t, God believes at t that p iff God knows at t that p iff p.[3]

From (6) and excluded middle it follows that either God knows that p or God knows that not-p, that is, God's knowledge and belief are negation-complete. Obviously, both are consistent.

Theorem. If X is omniscient, then there are no times t and t′ and no action type A such that at t, X intends to A at t′.

Proof. Suppose, for reductio, the following:

(a) At t, God intends to A at t′ (for times t, t′ and action-type A).

Then, by the description of intending given in Section II above, at a time no later than t, God adopts an intention to A at t′. So, since God fulfills the conditions of rationality and self-reflection mentioned in the antecedent of (4), then at some time t″ appropriately prior to t, God possesses the beliefs specified in the consequents of (3) and (5). Since God is omniscient, it follows from (a) and (6) that,

(b) At t″ God knows that at t he will intend to A at t′.

But, from (a), by (3) and (5), we also have,

(c) At t″, God believes that he himself does not yet believe that at t he will intend to A at t′.

and from this, by (6),

(d) At t″, God does not believe that at t he will intend to A at t′.

By (6), again,

(e) At t″, God does not know that at t he will intend to A at t′.

---

[3] This principle is set forth and defended in [8] pp. 265–6. The argument can be made even if God is only self-omniscient, i.e. knows every truth about itself, though 'p' would then be restricted to God's first-personal propositions.

which contradicts (b). Consequently, if God is omniscient, then the reductio suppposition (a) cannot be satisfied.

A complete grasp of the future must immobilize – a result devastating for any brand of theism which takes God to be an everlasting, omniscient agent. Far from being omnipotent, any being unfortunate (or fortunate?) enough to be omniscient is, of necessity, omni-impotent.[4]

East Carolina University,
Greenville, NC 27858, U.S.A.

REFERENCES

[1] M. Bratman, *Intentions, Plans, and Practical Reason* (Cambridge, MA: Harvard, 1987).
[2] D. Burrell, 'God's Eternity', *Faith and Philosophy* 1 (1984) 389–406.
[3] H-N. Castaneda, *Thinking and Doing* (Dordrecht: Reidel, 1976).
[4] H-N. Castaneda, 'Reference, Reality, and Perceptual Fields', *Proceedings and Addresses of the American Philosophical Association* 53 (1980) 763–822.
[5] D. Dennett, *Elbow Room* (Oxford: Clarendon Press, 1984).
[6] N. Denyer, *Time, Action and Necessity* (London: Duckworth, 1981).
[7] C. Ginet, 'Can The Will Be Caused?' *Philosophical Review* 71 (1962) 49–55.
[8] P. Grim, 'Some Neglected Problems of Omniscience', *American Philosophical Quarterly* 20 (1983) 265–6.
[9] S. Hampshire and H. L. A. Hart, 'Decision, Intention and Certainty', *Mind* 67 (1958).
[10] T. Kapitan, 'Deliberation and the Presumption of Open Alternatives', *The Philosophical Quarterly* 36 (1986) 230–51.
[11] A. Mele, 'Intention, Belief, and Intentional Action', *American Philosophical Quarterly* 26 (1989) 19–30.
[12] B. O'Shaughnessy, *The Will: A Dual Aspect Theory* (Cambridge: Cambridge University Press, 1980).
[13] J. Simons, 'Eternity, Omniscience, and Temporal Passage', *Review of Metaphysics* 42 (1989) 547–68.
[14] E. Stump and N. Kretzmann, 'Eternity', *Journal of Philosophy* 78 (1981) 429–58.
[15] R. Taylor, *Action and Purpose* (Engelwood Cliffs: Prentice-Hall, 1966).
[16] P. van Inwagen, *An Essay On Free Will* (Oxford: Clarendon Press, 1983).

---

[4] There are, of course, a number of possible theistic replies, the most significant of which endorses the Boethian image of God as 'timelessly eternal', thereby removing any doubt about God's uncertainty before formulating intentions. Recently, [13] and [14] have championed this view, though [2] expresses reservations, particularly, about the notion of 'atemporal duration' in [14]. Also, the eternality doctrine faces the problem of explaining how an eternal being can intend at all, since it could never have the required sense of doxastic 'openness'. Thus, what it 'does' would be more akin to instinctual or automatic doing, not intentional action. But if what we have said above is correct, namely that intentional action is impossible without intention-acquisition, and that this amounts to a change of state, then the literal image of an atemporal agent is an absurdity.

CLASSIFYING CONDITIONALS

By FRANK JACKSON

CONSIDER

(1) If Booth had not killed Lincoln, someone else would have
(2) If Booth does not kill Lincoln, someone else will
and
(3) If Booth did not kill Lincoln, someone else did.

Many writers agree that (1) is importantly different from (3). It is not just that we deny (1) and affirm (3) – a matter everyone agrees about – rather the explanation of why we deny (1) and affirm (3) is to be found in an important semantic difference between (1) and (3).[1] The details vary (needless to say), but there is considerable agreement that the key to the meaning of (1) lies in the fact that (1)'s truth conditions are to be given in terms of possible worlds in the style of Robert Stalnaker [13], or of David Lewis [11], or of some reasonable variant thereon; whereas the key to the meaning of (3) lies in the fact that the justified assertibility of (3) is given by the conditional probability of (3)'s consequent given its antecedent, or in terms of some reasonable variant thereon. (See e.g. Ernest Adams [1], [2].) This approach to (3)'s meaning is sometimes associated with the doctrine that (3) does not have truth conditions, and sometimes with the doctrine that (3) has the truth conditions of the material conditional.[2]

This paper takes for granted this general attitude to the relationship between (1) and (3). Our question is what then to say about (2). Should we group (2) with (1), or with (3)? If – as is common – we describe (1) as a subjunctive conditional and (3) as an indicative conditional, and go on to express the generalization of the idea that (1) and (3) are semantically distinct in an important way by saying that subjunctive and indicative conditionals are importantly semantically distinct, we presumably place (2) with (3). For (2) would commonly be classified as indicative. But this would be weak *qua* argument for classifying (2) with (3). (As has been widely noted, recently by Jonathan Bennett [3] and Michael Pendelbury [12].) The fact that (1) is subjunctive whereas (3) is indicative is far from the only difference between them which might reasonably be supposed to be the relevant one, particularly in view of the notorious elusiveness of the subjunctive/

---

[1] See e.g. Jonathan Bennett [3], Allan Gibbard [8], David Lewis [11] and Frank Jackson [10]. For a dissenting voice see Brian Ellis [6]. I follow Bennett in using Booth and Lincoln instead of Oswald and Kennedy. You are of course supposed to make a 'Warrenite' assumption about the example.

[2] For examples of the former response see Adams [2], and Dorothy Edgington [5]; for an example of the latter response see Jackson [10].

indicative distinction in English. Moreover, one of the main reasons for holding that (1) and (3) are semantically distinct is the noted point that we deny (1) and affirm (3); but if we go by denial and affirmation, then (2) most naturally goes with (1). As Brian Ellis in [6] and [7] has particularly emphasized, our reasons for denying (1) – the absence of a backup killer, for instance – are those which would justify denying (2) from a temporal perspective just prior to Booth shooting. What is more, suppose we acquired reason to affirm (1) by learning that Booth's wife was ready to fire from the stalls if he missed from the balcony; we would in that case hold that (2) was the right thing to say beforehand, and so going by affirmation and denial would in this case as well put (1) and (2) together.

However, the evidence from affirmation and denial for classing (1) and (2) together is not as strong as it might seem at first. There is an important difference between the data about affirmation and denial as it applies to separating (1) from (3), and as it applies to where to locate (2). From a given epistemological standpoint – roughly, that which we actually occupy – we deny (1) and affirm (3). That is the datum that most often motivates giving different accounts of (1) and (3). But it is not true that from a given epistemological standpoint we deny (2) while denying (1) and affirming (3). We deny (1) and affirm (3) from the standpoint we actually occupy after the event; but deny (2) from a standpoint we imagine ourselves to be occupying beforehand. There is no one epistemological standpoint from which we give different answers about (2) and (3). Hence, in the case of (2) and (3) it is an open possibility that the difference in regard to affirmation and denial can be explained by the difference in epistemological standpoints, thus allowing us to class (2) and (3) together, semantically speaking, and apart from (1).

The principal burden of this paper is that we should, semantically speaking, class (2) and (3) together, and apart from (1). The semantic line goes where the mood line goes, between (1) and (2), not between (2) and (3). After I have argued this, I will return to the question of providing an epistemological explanation of why we affirm (3) but from the imagined perspective before the event, deny (2).

I think that there are a number of reasons for classifying (2) with (3) (I give some of these reasons in [10], pp. 40f), but the case I will develop here turns on that provided by Allan Gibbard's Sly Pete example in [8]. I think that the significance of this much discussed example has been left rather obscure in the literature. Indeed, Bennett takes the example to provide reason for classifying (2) with (1).[3] I will argue, though, that once the lesson of the example is properly understood it provides a strong reason for classifying (2) with (3).

---

[3] See Bennett [3]. Michael Pendelbury [12] also holds (but for different reasons from Bennett's) that (2) should be classified with (1).

It will be useful to have labels for conditionals relevantly like (1), like (2), and like (3). I will call conditionals relevantly like (1) 'past subjunctives'; like (2) 'future indicatives'; and like (3) 'past indicatives'. I use these labels because they are convenient — you will know the conditionals I have in mind — and set to one side whether there are better labels to be had. Thus the main issue we will be concerned with can be expressed as how to locate semantically future indicative conditionals: with past subjunctives as Bennett holds, or with past indicative conditionals as I hold.[4]

I THE NATURAL APPROACH TO CONDITIONALS

My argument will be that although the natural approach to conditionals applies to past subjunctive conditionals, the Sly Pete example shows that it does not apply to past indicative conditionals and does not apply to future indicative conditionals. And, as will be clear when I describe the natural approach, this difference between past subjunctive on the one hand, and past and future indicative conditionals on the other, is most certainly a semantic one.

The natural approach is, as its name suggests, the natural one to take, the one we would like to take if we could. It has two clauses: (Support) and (Conditional non-contradiction).

(Support). On the face of it, acquiring information which justifies the assertion of a conditional is not different in kind from acquiring evidence which justifies the assertion of a non-conditional sentence. Reading the weather report justifies my saying that it will rain, reading the rules of cricket justifies my saying that if it rained heavily, the match was cancelled; learning about the stock exchange may justify my saying that a certain stock will go up, or it may justify my saying that if a certain stock goes up, certain other stocks will go down, and so on and so forth. By saying that there is, on the face of it, not a difference in kind I simply mean that both cases seem alike in essentially involving the acquisition of evidence that makes what is asserted more likely to be true. The information is support in the standard sense of making very probably true. (Support) is thus incompatible with the view that conditionals do not have truth conditions, only assertion conditions. But I think that it is important to appreciate that this deflationary position is one we need to be driven into; it is not one to embrace early on. For it makes a mystery of the sense in which information justifies, in the relevant sense, asserting a conditional. For instance, a lot of evidence that in the past, economic problems have been followed by election reverses for the government justifies saying that if the economy continues to go badly, the government will be in trouble at the next election. But this is not justification in the sense of making the saying morally justified, or pragmatically advisable, or in accord with good etiquette. What then is the relevant sense of justification? The obvious answer is that it is justification in the sense of making probably true.

---

[4] I am, of course, also disagreeing with Pendelbury [12]. But note that he labels (2) as well as (1) as 'subjunctive', see [12], p. 192. Hence, we come out in verbal agreement but remain in disagreement over the substantive question of (2)'s placement.

(Conditional non-contradiction). There is something very wrong with a given person saying both 'if *A* then *B*' and 'if *A* then not *B*', or at least there is when *A* is consistent. In tandem with this is the fact that if you say 'if *A* then *B*' and I say 'if *A* then not *B*', we count as disagreeing, or at least we do when *A* is consistent. (Reductio proofs in Logic and Mathematics show that when *A* is inconsistent the situation is different.) The obvious explanation of these facts is that 'if *A* then *B*' and 'if *A* then not *B*' are (logical) contraries when *A* is consistent. We will follow the practice of referring to this as (Conditional non-contradiction). (The name perhaps suggests the doctrine that, when *A* is consistent, not only are 'if *A* then *B*' and 'if *A* then not *B*' contraries, they are also contradictories in the sense that they must have opposite truth values. The argument that follows goes through on either interpretation, but we will work in terms of the weaker one.)

If a possible worlds semantics of one of the more familiar kinds applies to a given class of conditionals, then there is no problem about holding that the natural approach applies to that class. If 'if *A* then *B*' is true if and only if (something along the lines of) the closest *A*-worlds are all *B*-worlds, then 'if *A* then *B*' is (a) truth-valued, (b) supported by evidence in the standard sense because information garnered about how things actually are can make it more or less likely that nearby antecedent worlds are consequent worlds, and (c) inconsistent with 'if *A* then not *B*' when *A* is consistent, for the closest *A*-worlds cannot be both *B*-worlds and not-*B*-worlds. I think, in common with many, that the possible worlds semantics applies to past subjunctive conditionals. Hence, I take it that there is no problem about adopting the natural approach for past subjunctive conditionals, including of course (1).

The problem is with past and future indicative conditionals. The Sly Pete example shows that (unfortunately) the natural approach fails for them; although, as we will see, it takes a little work to see exactly how to argue the point.

II THE SLY PETE EXAMPLE AND SOME FALSE STARTS

There are a number of ways of setting out this example. We will work with the following version. It is in terms of past indicatives. We will discuss future indicatives shortly.

Sly Pete is playing poker. He has to decide between calling and folding. Informant *A* knows that Sly Pete is cheating, and that he is doing so by knowing what is in his opponent's hand. (Perhaps Sly Pete can see his opponent's hand in a mirror in the ceiling.) Informant *B* has surreptitiously observed both hands and knows that Sly Pete has the weaker hand. We can suppose that it is the last hand for the night in order to avoid the complication that Pete might want to lose the hand in order to encourage his opponent to bet higher later in the game. Informants *A* and *B* leave the room just before Pete makes his decision and report to you as follows. *A* says that if Pete called, he won. *B* says that if Pete called, he lost. You respond (correctly) by inferring that Pete did not call, he folded.

How does this example make trouble for the natural approach as applied to past indicative conditionals? Gibbard sees things this way. He points out that both informant *A* and informant *B* are fully justified in their assertions, are working from information properly so called, that is, which contains no mistakes, and are being sincere. From this he argues that 'we can see that neither is asserting anything false', and so that *if* 'both ... utterances express propositions, ... both express true propositions'.[5] To put matters in our terms, Gibbard is arguing that the (Support) part of the natural approach when combined with the fact that informant *A* and informant *B* are sincere and are proceeding in a fully justified manner from data that contains no mistakes would mean that what they each say is true, and so that (Conditional non-contradiction) fails. Hence, the natural approach is false, for its two clauses cannot be true together.

The trouble with this argument is that it is perfectly possible to proceed in a fully justified manner from information properly so called to a false conclusion. That is the point behind the familiar distinction between deductive and inductive arguments. In particular, it is perfectly possible for two different bodies of information relevant to a common subject matter to sustain incompatible conclusions about that subject matter. Smith knows that Petersen is a churchgoing Swede and that nearly all churchgoing Swedes are Protestants. She is entitled to infer with considerable confidence that Petersen is a Protestant. Jones knows that Petersen has Catholic parents and that nearly all churchgoers with Catholic parents are Catholic. She is entitled to infer with considerable confidence that Petersen is a Catholic. All this is compatible with Smith's information being information properly speaking, correct in every detail, and likewise for Jones's information. Of course, their two bodies of information are incomplete, but the same goes for the bodies of information available to informant *A* and to informant *B* in the Sly Pete example.

---

[5] Gibbard [8], p. 231. Helpful correspondence with Gibbard has convinced me that though my interpretation of Gibbard's argument is the best interpretation of the printed word, it is probably not the line of argument he intended, although I confess that the precise direction of the intended line of argument is still obscure to me.

There is, of course, a special feature of the Sly Pete example. In addition to the fact that both informants proceed justifiably from correct if incomplete information, there is the point that you, the hearer, are able to use their two utterances constructively to infer that Sly Pete did not call. This suggests that (a) 'if Sly Pete called, he won' conjoined with 'if Sly Pete called, he lost' entails that Sly Pete did not call, and that (b) it is rational to accept, in the circumstances as specified in the Sly Pete example, the conjunction of 'if Sly Pete called, he won' with 'if Sly Pete called, he lost'.

This way of construing the argument based on the Sly Pete example avoids the mistake of supposing that what comes rationally from what is the case must be true. It rests on the idea that if it is rational to accept 'if Sly Pete called, he won' together with 'if Sly Pete called, he lost', they can hardly be contraries.[6] The problem for this way of construing the argument is that it is not sufficiently obvious that our inference to 'Sly Pete did not call' depends *inter alia* on accepting the conjunction of the two conditionals. There is another hypothesis available to explain what is going on.

An important feature of the Sly Pete story is that you know the kind of information available to your two informants. You know that one informant knows that Sly Pete is cheating by knowing his opponent's hand, and that the other informant knows that Sly Pete's opponent has the stronger hand; and this fact plays a crucial role in warranting your inference from their two utterances to the conclusion that Sly Pete did not call. For suppose that you had had no idea of the evidence available to them, and that all you knew was that they were sincere and sensible. In that case you could have done nothing constructive with their two utterances, and in particular you could not have inferred that Sly Pete did not call. For suppose you cannot now remember whether you had your child inoculated against whooping cough, and one doctor says to you 'if you had your child inoculated against whooping cough, then you did the right thing', while another says to you 'if you had your child inoculated, you did the wrong thing'; you do not infer without further ado that you did not have your child inoculated against whooping cough. If in a sporting broadcast you hear Fred Stolle say that if Chris Evert won last night, she went one ahead in the head to head tally against Martina Navratilova, and then hear John Newcombe say that if Evert won, the tally is dead level, you do not take your high regard for the sincerity and good sense of Stolle and Newcombe as justifying the conclusion that Chris Evert lost. You simply infer that one of them has made a mistake.

---

[6] If *S* and *S** are sentences which are hard to comprehend fully, it can be rational to accept *S* along with *S** even when *S* and *S** are (unobviously) contraries. This is a point familiar to logicians and mathematics. But it is not plausible to think that our two sentences fall into this category.

Now, if it is crucial in the Sly Pete example that you know a good deal about your informant's evidence, there is an alternative hypothesis about how you reached your conclusion that Sly Pete did not call. We do not have to suppose that you inferred from the two conditional utterances taken together – a supposition which would indeed count against (Conditional non-contradiction) – we can instead suppose that you inferred from what you know of the evidence available to your informants. You know that informant *A* is using evidence that Sly Pete is cheating, informant *B* is using evidence that Sly Pete's hand is the weaker, and those two bits of evidence taken together are in themselves enough in the circumstances to warrant inferring that he did not call. We do not need to go via the informants' conditional utterances to explain how you were entitled to your conclusion.

In a variant on the Sly Pete example your informants' conditional utterances tell you something important about their respective bodies of evidence. In the variant you do not know initially that informant *A* knows that Sly Pete is cheating. You know merely that informant *A* knows whether or not Sly Pete is cheating, and knows nothing else relevant. Similarly, you do not know initially that informant *B* knows that Sly Pete has the weaker hand. You know merely that informant *B* knows who has the weaker hand, and knows nothing else relevant. In this variant case the conditional utterances of *A* and *B* do give you highly relevant information. When *A* says 'if Sly Pete called, he won', you learn that your informant knows that Sly Pete is cheating, for otherwise, given what you already know about *A*, *A* would not be justified in saying that if Sly Pete called, he won. When *B* says 'if Sly Pete called, he lost', you learn that your informant knows that Sly Pete's hand is the weaker one, for otherwise, given what you already know about *B*, *B* would not be justified in saying that if Sly Pete called, he lost. (We assume known sincerity and good sense in *A* and in *B*.) But even in this version of the Sly Pete example where your two informants' utterances play a major role in enabling you to conclude that Sly Pete did not call, we do not have to suppose that it is what they say as such that you are using. A believer in (Conditional non-contradiction) can insist that you are using what their two utterances tell you about their evidence, not the contents of the utterances themselves, to conclude that Sly Pete did not call. And this insistence would not be *ad hoc*. Believers in (Conditional non-contradiction) are entitled to make much of the point already emphasized that the two assertions of your informants, even if known to be sincere and sensible, in the absence of any information about the evidence that lies behind them, would not enable you to conclude that Sly Pete did not call. How so, if the evidence that backs their assertions is not the key?

III THE SLY PETE ARGUMENT AND A PRINCIPLE ABOUT EVIDENCE

Despite these failures, I think that it is possible to mount an argument from the Sly Pete example against the natural approach. I will refer to such an argument as a Sly Pete argument. It rests on a principle about evidence in favour of hypotheses which are logical contraries.

You might be tempted by the following principle about evidence in favour of contraries: if e_1 strongly supports H_1, e_2 strongly supports H_2, and H_1 and H_2 are contraries, then e_1 is unlikely given e_2, and e_2 is unlikely given e_1. But there is a counterexample to it. Set e_1 to be 'Petersen is a Swedish Catholic', e_2 to be 'Petersen is a Swede', H_1 to be 'Petersen is Catholic', and H_2 to be 'Peterson is not Catholic'. e_1 strongly supports H_1 (by entailing it), e_2 strongly supports H_2 (relative to a background which includes the fact that nearly all Swedes are Protestants), H_1 and H_2 are contraries, and yet it is false that e_2 is improbable given e_1. In fact e_1 entails e_2. It is, however, true that e_1 is improbable given e_2 (relative to the background evidence), and indeed the following principle is valid.

> (P) If e_1 is strong evidence for H_1, e_2 is strong evidence for H_2, and H_1 and H_2 are contraries, then e_1 is improbable given e_2 or e_2 is improbable given e_1 (or both).

More precisely, we can specify a value such that if $\Pr(H_1/e_1)$ and $\Pr(H_2/e_2)$ both exceed it, one or both of $\Pr(e_1/e_2)$ and $\Pr(e_2/e_1)$ are <0.5.[7]

It is by reference to (P) that we show that the Sly Pete example refutes the natural approach to conditionals as applied to past indicatives. Informant *A*'s information is that Sly Pete knows what is in his opponent's hand. This information along with such facts as that Sly Pete wants not to lose money on this hand and understands the rules of poker, very strongly warrants *A*'s assertion that if Sly Pete called, he won. Similarly, informant *B*'s information that Sly Pete has the weaker hand, in the appropriate circumstances very strongly warrants *B*'s assertion that if Sly Pete called, he lost. However, Sly Pete's knowing his opponent's hand need not count against Sly Pete's having the weaker hand. Similarly, Sly Pete's having the weaker hand need not count against his knowing his opponent's hand. Typically, 'Sly Pete knows his opponent's hand' and 'Sly Pete has the weaker hand' are evidentially neutral with respect to each other. In the example as originally described by Gibbard, Sly Pete has an accomplice who signals the contents of the opponent's hand to Pete. We imagined that Pete can see his opponent's hand reflected in a mirror in the ceiling. Typically, the successful operation of systems of these kinds are probabilistically quite independent of which hand is the stronger of the two.

---

[7] One way to get the result is to start with the special case where H_1 and H_2 are contradictories so that, for instance, $\Pr(e_1/e_2)<0.5$ iff $\Pr(e_1 e_2 H_1)+\Pr(e_1 e_2 H_2)<\Pr(e_2 H_1 \sim e_1)+\Pr(e_2 H_2 \sim e_1)$.

The same point can be made in terms of other examples. Information that a certain motion put at a certain meeting was a bad one would strongly warrant asserting 'if Fred voted for the motion, he did the wrong thing'. Information that Fred has excellent judgment in such matters would strongly warrant asserting 'if Fred voted for the motion, he did the right thing'. However, the fact that a certain motion was bad does in itself not count against (or for) the quality of Fred's judgment; and, conversely, Fred's excellent judgment does not in itself count against (or for) the motion being bad. We must conclude, therefore, that the Sly Pete example shows via (P) that the natural approach fails for past indicative conditionals. It cannot simultaneously be the case that (a) the two bodies of information that very strongly warrant the assertion, respectively, of 'if Sly Pete called, he won' and 'if Sly Pete called, he lost', are evidence in the standard sense of making the conditional whose assertion is warranted highly probably true, and (b) the conditionals are logical contraries. (a) and (b) cannot be true together because the two bodies of information may each be not improbable given the other. It follows that (Support) and (Conditional non-contradiction) clash in the case of past indicatives. But as we saw (or rather noted had been widely granted), the natural theory is true for past subjunctives, so we have the result that past indicative conditionals are semantically distinct from past subjunctive conditionals.

By itself this result is no advance. We *started* by noting the semantic distinctness of (1) from (3). We are now though in a position to argue that (2), the future indicative, should be classed with (3), the past indicative, rather than with (1), the past subjunctive. We can, now say with some precision how a Sly Pete argument can show that the natural approach to conditionals fails for conditionals of some specified class. It does so by describing a case where there are two bodies of evidence e_1 and e_2, and a pair of conditionals 'if A then B' and 'if A then not B' of the class in question with A consistent, such that against some given background: (a) e_1 by itself strongly warrants (as highly as you care to make it by varying the background) asserting 'if A then B', (b) e_2 by itself strongly warrants asserting 'if A then not B', and yet (c) e_1 is not improbable given e_2, and e_2 is not improbable given e_1.

A Sly Pete argument can be mounted for future indicative conditionals in much the same way as for past indicative conditionals.[8] Although the Sly Pete example is commonly described in the past tense, this is an accidental feature of the example. We imagined your two informants reporting to you after Sly Pete's decision, but we could equally have imagined them to be reporting just beforehand. Informant *A* knows that Sly Pete knows his opponent's hand and says to you, with full justification, 'if Sly Pete calls, he will win'. Informant *B* knows that Sly Pete's hand is the weaker and says, with full justification, 'if Sly Pete calls, he will lose'. And, as before, *A*'s information does not count against *B*'s, and *B*'s information does not count against *A*'s. Thus the case against the natural theory applying to future indicative conditionals is as strong as that against its applying to past indicative conditionals.

IV Sly Pete and Past Subjunctives

An obvious question to ask at this point is whether a Sly Pete argument is possible for past subjunctive conditionals. It had better not be, given our earlier presumption that the natural approach applies to past subjunctives.

I think, however, that we can see that a Sly Pete argument is not possible for past subjunctives. If we take the Sly Pete example and replace the pair of past indicative conditionals by the corresponding pair of past subjunctive conditionals, we get 'if Sly Pete had called, he would have won' and 'if Sly Pete had called, he would have lost'. Now the second of these conditionals is supported by the information that warrants the assertion of the corresponding past indicative (or the corresponding future indicative, if it comes to that). The information that Sly Pete had the weaker hand justifies saying both that if he called, he lost, and that if he had called, he would have lost. However, the first of these past subjunctives — 'if Sly Pete had called, he would have won' — is not warranted by the information about his cheating that warrants the assertion of the corresponding past indicative.

To see this consider the matter from the perspective of informant *A* who is in possession of this information, that is, the information that Sly Pete is cheating by knowing his opponent's hand. We noted before that this information is neutral concerning who has the stronger hand, so let us suppose that *A* gives equal credence to the hypothesis that Sly Pete's hand is stronger, and to the alternative hypothesis that it is weaker than his opponent's. What *A* can be sure about though is that Sly Pete will do the right, in the sense of the most rewarding, thing. That is his return for cheating. But what is the right thing for Sly Pete to do? The answer depends on whether or not he has the stronger hand. If he has the stronger hand, the right thing to do is to call; if he has the weaker hand, the right thing to do is to fold. Hence, informant *A* gives a 50 per cent credence to calling being the right action for Sly Pete, and 50 per cent to folding being the right action for Sly Pete. But what makes calling the right action for Sly Pete? Well, he had two options — to call or to fold, and calling is best if it had or would have had the better consequences of the two. But that is the case precisely if had Sly Pete called, he would have won. That is to say, the right credence for informant *A* to give 'if Sly Pete had called, he would have won' is the same as the credence he gives to calling being the right thing for Sly Pete to have done, which is 50 per cent. The upshot is that although informant *A*'s information that Pete is cheating makes 'if Sly Pete called, he won' very highly assertible, it leaves the credence of the corresponding past subjunctive unchanged at 50 per cent.[9] In brief, informant *A*'s information that Pete is cheating justifies asserting '*if* Sly Pete did not call, then had he called, he would have lost', but does not justify asserting outright either that had he called he would have lost, or that had he called he would have won.

---

[8] As Gibbard [8] p. 228 notes, though he also notes some complications that space precludes our pursuing.

V WHEREIN LIES THE SEMANTIC DIFFERENCE?

Our argument has been that *there* is a semantic difference between past subjunctive conditionals on the one hand, and past and future indicative conditionals on the other: the natural approach applies to the former but not to the latter. But our argument has been silent on wherein lies the difference, for we have not addressed the question of precisely *where* the natural approach fails for the past and future indicative conditionals. There would appear to be three live possibilities. I will not try to adjudicate between them here. I mention them to give a sense of the possibilities opened up by the Sly Pete argument.

You might hold that (Support) fails. Information that (strongly) justifies asserting past and future indicative conditionals does not make them (highly) probably true. The most familiar version of this response holds that past and future indicative conditionals are not truth-valued to start with, and so that there is no question of anything making them probably true. (See e.g. Adams [2], Gibbard [9], and Edgington [5].)

Secondly, you might hold that it is (Conditional non-contradiction) which fails. Even when *A* is consistent, 'if *A* then *B*' and 'if *A* then not-*B*' can be true together. The most familiar version of this response gives past and future indicative conditionals the truth conditions of the material conditional.

---

[9] This point has of course played a central role in recent discussions of decision theory, see, e.g., Allan Gibbard and William L. Harper [9].

Finally, you might hold that what is perhaps best described as a presupposition of (Conditional non-contradiction) fails. For you might hold that the truth conditions of past and future indicative conditionals are a function of the context of assertion, that is to say, that the proposition expressed by 'if *A* then *B*' varies with that context. More particularly, you might hold that (Conditional non-contradiction) is valid in the following sense: relative to a given context, the proposition expressed by 'if *A* then *B*' and the proposition expressed by 'if *A* then not-*B*' are contraries provided *A* is consistent. This is why it is wrong for a given person at a given time to assert both together. However, what may happen is that the proposition expressed by 'if *A* then *B*' relative to one context is consistent with the proposition expressed by 'if *A* then not-*B*' relative to some other context. (See e.g. Stalnaker [14].) And you might hold that exactly this happens in the Sly Pete example: that is to say, you might hold that the proposition that *A* expresses by saying 'if Sly Pete called, he won' is consistent with the proposition that *B* expresses by saying 'if Sly Pete called, he lost' though it is not consistent with the proposition that *A* would have expressed had *A* used *B*'s words.

(Of course it just might be that the way in which the standard approach fails for future indicative conditionals is different from the way it fails for past indicative conditionals. This is a consequence of the fact that the standard approach has *two* clauses.)

VI EPISTEMOLOGY AND THE DATA ABOUT AFFIRMATION AND DENIAL

We noted at the very beginning that if we went by affirmation and denial, we would class (1) and (2) together, and so separate them from (3), and I acknowledged an obligation *qua* proponent of classing (2) with (3) semantically, to give an epistemological explanation of the fact that we deny (1), would beforehand deny (2), and affirm (3), and I noted the abstract possibility of doing this. I now want to make this abstract possibility concrete by describing a situation in which we would deny (1) and affirm both (2) and (3), a case, that is, where going by affirmation and denial places (2) and (3) together. It will be apparent that the difference between this situation and our actual one is an epistemological one.

Our justified beliefs about what will happen in the future often depend heavily on our opinions about what the past and present will give rise to. My beliefs about who will win the next election depend heavily on my opinions about the likely consequences of present and past conditions. By contrast, my opinion — indeed, knowledge — about who won last year's election is relatively independent of the more or less speculative views I have about what caused that electoral victory. Similarly, my view beforehand about Lincoln's being killed would normally depend on my view about how it might happen that he comes to be killed. But we can have exceptions to this norm. I may expect Lincoln to be killed by being killed by Booth working alone, and yet be more confident that he will be killed by someone or other working alone than I am that it will be Booth. Perhaps someone I know to be very knowledgeable about the plot has assured me that Lincoln will be assassinated by someone acting alone, but she remained silent on who it would be. Booth is just my educated guess.

What do I say beforehand? I affirm (2), that if Booth does not kill Lincoln, someone else will. I expect it to be Booth, but be I right or wrong about that part of the story, what I am confident of is that it will be someone or other, and so, if not Booth, then someone else. What do I say afterward when I learn that Lincoln was indeed killed as predicted, that is, by someone working alone, and that it was almost certainly, as I had anticipated, Booth? Clearly I do not affirm that if Booth had not killed Lincoln, someone else would have. Booth acted alone. If he had not killed Lincoln, the plot would have failed and Lincoln would not have been killed. So I deny (1). Finally, provided that although I am confident that it was Booth who killed Lincoln, I am even more confident that Lincoln was killed, I will affirm (3), namely, that if Booth did not kill Lincoln, someone else did.

We have, therefore, concrete reason to hold that the fact that in the normal case we would, going by affirmation and denial, put (2) with (1), and not with (3) has an epistemological explanation not a semantic one, a reason to do with the difference in the epistemic status that normally obtains between opinions about the future and the past. For if we change that status in the appropriate way, going by affirmation and denial puts (2) with (3) not (1).

My case for separating semantically past subjunctive conditionals on the one hand from past indicative conditionals and future indicative conditionals on the other is now before you. Is it theoretically desirable to describe this by saying that the, or anyway an, important line between conditionals is between indicative conditionals and subjunctive conditionals? This depends on what should be said about future subjunctive conditionals, about which we have said nothing, and on the more general question of the viability in English grammar of dividing indicative from subjunctive moods. Here I am content to insist that in respect to one major feature the line has (1) on one side and (2) and (3) on the other.[10]

Research School of Social Sciences,
Australian National University, ACT 2601

---

[10] I am indebted to too many helpful comments on earlier versions to list, but must mention Lloyd Humberstone, David Lewis, and especially Peter Menzies who gave me invaluable help with (P).

REFERENCES

[1] Ernest Adams, 'The Logic of Conditionals', *Inquiry* 8 (1965) 166–97.

[2] Ernest Adams, *The Logic of Conditionals* (Dordrecht: Reidel, 1975).

[3] Jonathan Bennett, 'Farewell to the Phlogiston Theory of Conditionals', *Mind* 97 (1988) 509–27.

[4] V. H. Dudman, 'Parsing "If" – Sentences', *Analysis* 44 (1984) 145–53.

[5] Dorothy Edgington, 'Do Conditionals have Truth Conditions?', *Critica* 98 (1986) 3–30.

[6] Brian Ellis, 'A Unified Theory of Conditionals', *Journal of Philosophical Logic* 7 (1978) 107–24.

[7] Brian Ellis, 'Two Theories of Indicative Conditionals', *Australasian Journal of Philosophy* 62 (1984) 50–66.

[8] Allan Gibbard, 'Two Recent Theories of Conditionals', in *Ifs*, edited by W. L. Harper *et al.* (Dordrecht: Reidel, 1981), 211–47.

[9] Allan Gibbard and William L. Harper, 'Counterfactuals and Two Kinds of Expected Utility', reprinted in *Ifs*, edited by W. L. Harper *et al.* (Dordrecht: Reidel, 1981), 153–90.

[10] Frank Jackson, *Conditionals* (Oxford: Blackwell, 1987).

[11] David Lewis, *Counterfactuals* (Oxford: Blackwell, 1973).

[12] Michael Pendelbury, 'The Projection Strategy and the Truth Conditions of Conditional Statements', *Mind* 98 (1989) 179–205.

[13] Robert Stalnaker, 'A Theory of Conditionals', in *Studies in Logical Theory, American Philosophical Quarterly*, monograph series, no. 2, edited by N. Rescher (Oxford: Blackwell, 1968), 98–112.

[14] Robert Stalnaker, *Inquiry* (Cambridge, Mass.: MIT Press, 1985).

Vol. 50 No. 3 (New Series No. 227) June 1990

ANALYSIS

Edited by
PETER SMITH

Success semantics J. T. WHYTE
Vague identity yet again HAROLD W. NOONAN
Contingent identity in counterpart theory MURALI RAMACHANDRAN
Counterparts, logic and metaphysics: reply to Ramachandran GRAEME FORBES
Unsuccessful revisions of CCT MURALI RAMACHANDRAN
Truth theoretical semantics and ambiguity BRENDAN S. GILLON
Fitch and intuitionistic knowability PHILIP PERCIVAL
Two uses of 'know' D. S. CLARKE, JR.
Intending and motivation JOSEPH MENDOLA
Intending and motivation: a rejoinder ALFRED R. MELE
When do we deceive others? ANNETTE BARNES
Rationality in conversation and neutrality in politics GOVERT DEN HARTOGH
The divestiture puzzle dissolved ROGER A. SHINER
The pragmatic problem of induction: reply to Gower and Bamford
JOHN WATKINS

BASIL BLACKWELL · OXFORD AND CAMBRIDGE, MA

ISSN 0003-2638

ANALYSIS 50.3 JUNE 1990

SUCCESS SEMANTICS

By J. T. WHYTE

I TRUTH: WHAT IT IS AND WHY WE WANT IT

TRUTH is the property we most want our beliefs to possess: and not because truth is beauty and we want our beliefs to be beautiful; nor for any other especially noble reason. We want our beliefs to be true simply because, when they are, we get what we want by acting on them. The truth of beliefs explains the success of the actions they cause. Although obvious, this is the most interesting fact about truth, because it tells us not only why we want our beliefs to be true but also what truth is. Truth just *is* the property of a belief that suffices for your getting what you want when you act on it.

And, of course, we already know what property that is. It's the property of being a belief that things are thus-and-so when they *are* thus-and-so. For clearly, you will get what you want if you act on a belief that represents things as being as they actually are. Plumpton will succeed in becoming thin by acting on his belief (B1) that dieting will make him thin, just in case dieting *will* make him thin.

So, thinking of truth in this way simply delivers a *redundancy* conception, according to which

(RT) a belief B is true iff $(\exists p)$(B is the belief that p & p).

Which is fine, as far as it goes, but that's not far enough for those of us who think there must be more to truth than (RT) gives us. Even if the redundancy conception is right we still need to say what makes any belief the belief it is. What, for example, makes Plumpton's belief B1 the belief that dieting will make him thin and thus true if and only if dieting will make him thin? The substantial part of a redundancy account of truth lies in saying what gives beliefs their truth conditions.

And this is where the connection between truth and success becomes genuinely interesting. For taking truth to be the guarantor of successful action delivers, not only the redundancy conception, but also an account of truth conditions. For suppose we pretend that we don't know the truth condition of Plumpton's belief B1, but do know that combined with the desire to be thin, B1 makes Plumpton diet. Now beliefs, we've said, are true just in case they suffice to get you what you want when you act on them.

So, B1's truth condition is what suffices for Plumpton's action getting him what he wants: namely, that Plumpton's dieting *will* make him thin. *That* is B1's truth condition.

Thus we arrive at the following account of truth conditions, which I'll call (R), after Ramsey, who thought of it first (see F. P. Ramsey 'Facts and Propositions' in his *Foundations: Essays in Philosophy, Logic, Mathematics and Economics* edited by D. H. Mellor, Routledge & Kegan Paul, 1978):

> (R) A belief's truth condition is that which guarantees the fulfilment of any desire by the action which that belief and desire would combine to cause.

Since (R) defines truth conditions in terms of the content of desires (their fulfilment conditions), it doesn't itself reduce truth conditions to anything non-semantic. And clearly if we want such a reduction, and want to stick with (R), then we will have to give a 'truth independent' account of the content of desires. That can be done, but isn't what I'm concerned to show here. The aim of this paper is simply to show that, taking the content of desires for granted, (R) gives the right answers regarding the truth conditions of beliefs.

II TRUTH CONDITIONS FOR INDIVIDUAL BELIEFS

You might think that (R) will work only for beliefs, like Plumpton's B1, which are about means to ends. Why should (R) work for beliefs whose content isn't instrumental in this sense, and which therefore don't, on their own, combine with desires to cause any actions?

The answer is that *all* beliefs are instrumental in the required sense. Certainly, not all are made true by facts about what will happen if something is done (instrumental facts), but that doesn't matter. All (R) needs – and gets – is that all beliefs have distinctive causal properties: having a belief systematically affects what you do, or at least what you would do if relevant circumstances arose.

Indeed, beliefs can be identified, without assuming anything about truth conditions, with their causal properties by the following conjunctive characterization (call it CAB – short for 'causal account of beliefs'):

> (CAB) Bi is the state-type that (if combined with the desire D1 and beliefs Bj, Bk, . . .) would cause the action A1 and (if combined with D2 and Bn, Bm . . .) would cause A2 and . . . [and so on, listing all the actions that Bi would cause if combined with other beliefs and desires.]

CAB identifies beliefs dispositionally. Being dispositions is what makes all belief states instrumental, even ones about how many toes you have on your right foot. All beliefs will cause you to act in a certain way given certain other beliefs and desires. Plumpton's B1 is more obviously instrumental only because it can be characterized causally without referring to any other beliefs: because it can combine with a desire to cause an action on its own.

But how can (R) give the right truth conditions for beliefs that need other beliefs (as well as desires) to make them cause actions? Here's how. Suppose that Plumpton (who has given up his diet) now wants some chocolate and believes (B2) that there is chocolate in the fridge. B2 won't on its own make Plumpton act on his desire for chocolate by going to the fridge. He won't, for example, go to the fridge if he thinks that it's locked. To get Plumpton to the fridge B2 needs to combine with other beliefs: if not beliefs about the fridge not being locked etc., at least beliefs whose truth entails such things. Suppose then, that B2 does combine with some other such beliefs, say B3 and B4, to make Plumpton go to the fridge (call this action A2). What can (R) tell us about B2's truth condition?

What it tells us directly is the truth condition for the conjunction of beliefs B2 & B3 & B4. This is a condition (or conjunction of conditions) sufficient for Plumpton's getting chocolate by going to the fridge: there being chocolate in the fridge, no impediments to Plumpton's making it to the kitchen, and so on. But we want an account of the truth conditions of individual beliefs, not just of conjunctions of them. But that isn't hard to extract, via the fact that any one belief will combine with many other beliefs and desires to cause equally many different actions.

Let's stick with Plumpton's belief B2. How can we pick out the part of the A2's success condition to which B2 answers for its truth (namely, there being chocolate in the fridge)? Very simply. Take an action caused by B2, but not B3 and B4, and see what that action's success condition shares with A2's success condition. Suppose, for example, that Plumpton wants to go and buy any confectionery he doesn't already have in his fridge. B2 will play its role here: it will stop Plumpton from buying chocolate. But B3 and B4 — the beliefs about Plumpton's kitchen (etc.) that were required to get into it — will have no effect on Plumpton's purchases; and their truth will have no effect on whether those purchases succeed. But B2's truth will affect that; since if it is false Plumpton will not buy some confectionery that isn't in his fridge. And so, in short, what makes B2 the belief that there is chocolate in the fridge is that wherever B2 is part of the cause of an action A, part of A's success condition is that there is chocolate in the fridge.

So far so good. But what about conditions that are part of every action's success condition? Or those which are part of the success conditions of all actions caused by beliefs about certain things: for example, the condition of the earth's gravitational field, which is part of the success condition of all actions caused by beliefs about masses on the earth?

The problem these ever-present success conditions presents is obvious. Wherever any belief (or belief of a certain type) B causes an action A, these conditions will be part of A's success condition; but they need not be part of B's truth condition. For example, almost no action would succeed if you exploded whenever you tried to perform it. But clearly that doesn't make the truth conditions of our beliefs include our not exploding. B2 is not Plumpton's belief that there is chocolate in the fridge *and* that he won't explode if he does anything. So perhaps (R) is in trouble. Perhaps it is committed to the ludicrous view that 'stability conditions' (conditions required for the success of all actions) are part of the truth conditional content of all beliefs.

Not so. This trouble only seems to arise because we are overly concerned with how things actually are. Of course, the *actual* success of our actions depends, as a matter of fact, on certain regularities. But (R) is not restricted to *actual* success. Recall how we removed the fridge not being locked from the truth condition of Plumpton's belief B2: by considering hypothetical actions caused by B2 whose success would be independent of whether the fridge was locked. And as for the fridge not being locked, so for more general stability conditions. We need only consider what would remain part of the success conditions of actions partly caused by B2 if the stability conditions were different.

For example, terrestrial gravity no doubt plays a part in the success of all Plumpton's actions involving fridges. But that doesn't commit (R) to making all Plumpton's beliefs about fridges (including B2) also beliefs about terrestrial gravity. For if there were no gravity, and thus no true beliefs about its obtaining, B2 could, and sometimes would, still be part of the cause of successful actions. Gravity is no part of B2's truth condition because even though it is part of the success condition of every action it *does* cause, it is not part of the success condition of every action it *would* cause if things were different.

On the other hand, no matter how things differed, so long as B2's causal dispositions remained fixed, the success condition of any action it caused would include chocolate being in the fridge. For example, if Plumpton were in outer space and wanted chocolate, then combined with some other beliefs, B2 could make him perform an action whose success required the *absence* of gravity: such as wafting gently towards the fridge hovering at the ceiling of his space-ship. But however he were to get to the fridge, there would still have to be chocolate in it for Plumpton to get what he wants.

So, to single out a belief's truth condition we only need counterfactuals about what would suffice to make all the actions it would cause succeed if only things were different, provided, of course, its causal dispositions remain fixed.

But doesn't this reveal at least one kind of regularity which we *can't* peel away from a belief's truth condition: namely, the very causal regularities of CAB that we used to characterize the belief in the first place? For they, at least, will be present wherever the belief they characterize is. But they aren't in its truth condition. How can we stop (R) making them so?

Again, quite simply. First we note that no such causal regularities will be part of the success condition of *all* the actions caused by the beliefs they define. For the success condition of an action is what suffices for it to fulfil the desire that caused it. And that will *very rarely* include the regularities that define the beliefs that also caused it. What those regularities explain is the actions performed, not why, once performed, those actions are successful.

Consider again Plumpton's belief B2 and his expedition to the fridge. The regularities that link B2, B3, B4 and D2 to Plumpton's going to the fridge do indeed partly explain that action of his; but they are irrelevant to its success once it's done. So even though beliefs must always be accompanied by the casual conditions they depend on for their identity, those conditions aren't what the beliefs are about, because they aren't always part of the success conditions of the actions caused by the beliefs they identify.

But there is another objection to consider. This appeals to the fact that the counterfactual strategy for singling out a belief's truth condition requires assuming the world to be different and seeing what would suffice for the success of the actions it would then cause. And if the world were different in some such ways, in particular in some of its regularities, then it would differ in many other ways, including other regularities. But if lots of our regularities were different then, since we have defined beliefs (in CAB) by way of the regularities that govern them, our beliefs would have to differ too. And I can't talk about what B2 would cause if there were differences in the regularities on which B2 depends for its very existence.

But there's no real problem here either: because we don't need to consider situations which vary so radically from actuality that our beliefs couldn't exist in them. In particular, we need not deny all the regularities that we want to exclude from the truth conditions of our beliefs all at once. We can peel them away one at a time, by considering, for each such regularity, what would suffice for the success of actions caused by the belief in question if only that one regularity were different or irrelevant. And by also keeping other differences to a minimum (as you always do with counterfactuals) there will be no problem in talking about what actual beliefs would cause if things weren't as they actually are.

But won't even this cautious approach sometimes require us to consider what would happen in situations extremely unlike reality: for example, situations where acting *does* make us explode and in which case the regularities of CAB, which characterize our beliefs, will surely not hold?

No it won't. For to get something out of one belief's truth condition, it is good enough to get it into another's, as follows. We've already noted that the success conditions of the actions a belief causes generally include conditions that aren't part of its truth condition, and that this presents no problem because such beliefs only cause actions when combined with other beliefs. For example, the fridge's not being locked is not in B2's truth condition because it is part of the truth conditions of B3 and B4, the other beliefs with which B2 combined to cause A2. Similarly, that acting won't make you explode, needn't be in the truth condition of B2 because it can be in – indeed can be – the truth condition of another belief, B5, without which you would never act at all.

But how are we to distinguish the truth-conditional content of B2 from that of B5, given that B2 only causes actions when combined with B5?

Easily. For even if B2 cannot cause action independently of B5, the converse doesn't hold: B5, but not B2, is part of the cause of most actions. And that is what makes it, and not B2, the belief that acting won't make you explode: that acting won't make you explode is part of the success condition of all the actions that B5 would cause. And in order to single out *that* truth condition we needn't consider situations in which acting does make you explode, but merely situations in which the other beliefs, with which B5 combines to make you act, aren't true; such as there being no chocolate in the fridge. And nothing in those situations need threaten the regularities on which the existence of any of our beliefs depends.

There's an easier and perhaps less artificial way to show that employing (R) doesn't make us consider overly bizarre situations. Beliefs about things as peculiar as your not exploding seem very rarely to play any part in causing actions. We act all the time, but we hardly ever (at least consciously) believe that acting won't make us explode. In addition to B2 and the desire for chocolate, it's probably enough to get Plumpton to the kitchen that he believes (B6) that there's nothing to stop him getting there; the truth of which *implies* that acting won't make him explode. By implying B5, B6 can make B5's causal contribution without Plumpton having to have B5. Now as we know, in various conjunctions of beliefs and desires, B2 can make Plumpton perform many different actions. But unlike B5, B6 won't often be one of the conjuncts. There will, however, almost always be one belief that implies that he won't explode.

It is now clear how we can avoid having to consider situations which threaten CAB. Since many different beliefs imply B5, to single out B2's truth condition I need never consider what would happen if B5 were false (if acting did make you explode), but only what would happen if the particular beliefs that imply B5 were false. And, unlike B5, to be false (or irrelevant to the performance and success of actions) such beliefs don't need to be held in situations where acting will make you explode. The falsity of B6, for example, requires nothing more exotic than the kitchen door being locked.

Enough has been said, I think, to show how (R) can single out truth conditions for individual beliefs. Before moving on, however, we should note that this strategy will single out even more fine grained contents: namely, the references of individual concepts. Suppose we not only have truth conditions for all our beliefs but also that their causal inter-relations given them a sub-propositional structure. Then the referents of concepts – the mental correlates of words – can be singled out in the same way as the truth conditions of whole beliefs. If an entity E occurs in the truth conditions of all beliefs that include a concept C, then C refers to E. For example, if one of B2's sub-propositional components, C1, is also a part of other beliefs, all of whose truth conditions involve Plumpton's fridge but have nothing else in common, then C1 refers to Plumpton's fridge.

That completes my brief explication of (R) and how it works. Of course, I can't here deal with many of the issues (such as semantic reduction) raised by (R), but I will say something about two possible grounds of objection to it: the role validity plays in (R), and the way we can be caused to act by partial beliefs.

III Does (R) Need Validity?

You might think that (R) gives a circular account of truth because it requires beliefs to cause actions by way of valid inferences. After all, if Plumpton invalidly inferred from B2 that if he wants chocolate he should go to the washing machine, (R) would give B2 the wrong truth condition: namely, that there is chocolate in the washing machine. So perhaps (R) needs to be augmented with a validity stipulation:

> (R′) A belief's truth condition is that which guarantees the fulfilment of any desire by the action which that belief and desire would combine to cause by way of valid inferences.

But validity is defined in terms of truth: valid inferences are those that are necessarily truth preserving. So adding a validity stipulation would make (R) give a circular account of truth. But there's no problem here because (R) doesn't actually need to appeal to validity. All (R) needs is some psycho-physical laws, such as those described by CAB. Since they are what determine the truth conditions of beliefs, they will indeed characterize valid inferences between intentional states: but that is a consequence, not an assumption, of (R).

But how then can people make invalid inferences, and fail to believe all the consequences of their beliefs? Of course, they sometimes do, but this is perfectly compatible with (R): like other dispositional or functional states, beliefs needn't always have the effects that characterize them. For one thing, most of the laws described in CAB will be probabilistic, and people needn't always do what a law says they only usually do. For another, dispositional states are typically states of systems governed by many laws besides the laws which characterize those states. All sorts of things could happen to your body to stop you acting in accordance with the laws that characterize your mind. But that doesn't refute the dispositional account of states of your mind. Being physically unable to do sums, because the activity of your brain makes you sleep, doesn't stop the conclusion you would arrive at if you had a better brain from partly fixing the content of that calculating state that in fact only sends you to sleep.

So (R) doesn't imply that mistakes are impossible. But you may still think that (R) only looks plausible if the laws of CAB do characterize valid inferences between intentional states. Otherwise why should a belief's truth condition be linked to the success conditions of the actions it causes? Why shouldn't (R) let B2 send Plumpton to the washing machine for his chocolate if (R) doesn't presuppose validity? The answer is that if B2 did that (and did so in accordance with psycho-physical law), it wouldn't be the belief it is: it wouldn't be the belief that there is chocolate in the fridge; it would be the belief that there is chocolate in the washing machine.

IV PARTIAL BELIEF

So much for validity. What about partial belief? We do, after all, often act on beliefs that we hold with less than absolute certainty. But then the truth of those partial beliefs doesn't guarantee the success of some of the actions they cause. Suppose, for example, that Plumpton's degree of belief in B2 is only 50%. That partial belief could combine with, say, a desire not to expend energy on a fruitless trip to the fridge, to cause Plumpton to stay seated in front of the television. But it isn't part of the success condition of Plumpton's sitting either that there is chocolate in the fridge or that there isn't. How can (R) allow truth and success conditions to come apart as they clearly do in these cases?

Easily: by not applying to them. A belief's truth condition is that which suffices for the success of the actions it would cause *if it were a full belief*. It doesn't matter to (R) that beliefs are often only partial. Since coming to England, for example, I have never fully believed (B7) that tomorrow will be sunny. And my partially believing B7 has indeed made me do things whose success condition doesn't include the next day being sunny. But B7 is still the belief that tomorrow will be sunny, because tomorrow's being sunny is part of the success condition of all the actions it would cause if I believed it fully; even though, not being that stupid, I never will.

In short, we don't need the whole of modern decision theory to fix the truth conditions of our beliefs. We only need the limiting case, which tells us how full beliefs make us act. That will suffice to fix their truth conditions; and we can then use the rest of decision theory to tell us how beliefs with those truth conditions will make us act when they are only partial beliefs. That's another story; but not one that need concern us here, since far from contradicting (R), it ultimately presupposes it.[1]

St John's College,
Cambridge CB2 1TP

---

[1] I am grateful to Hugh Mellor for helpful comments.

VAGUE IDENTITY YET AGAIN

By HAROLD W. NOONAN

GARETH EVANS'S article on vague identity [1] has been the subject of many criticisms. Despite these criticisms, however, I believe that fundamentally (despite some confused remarks on Evans's part about S5 and over-brevity in his explanation of his determinacy operator 'Δ'), Evans's attack on the notion of indefinite identity is sound. In what follows I first present what seems to me the most powerful version of Evans's argument and then respond to the recent criticisms of Over [6], Garrett [2] and Johnsen [3]. In doing so I attempt to make it clear that there is, in essence, only one reply to Evans available to the defender of indefinite identity, and that a most uninviting one, namely the defence of indefinite identity as a kind of *relative identity* (which I call 'mere indistinguishability in non-delta properties') and the rejection of a principle I shall refer to as 'the principle of the Diversity of the Definitely Dissimilar'.

As David Lewis has stressed in his [4], Evans is not against the idea that there can be identity *statements* which are indeterminate in truth-value. Since there are obvious examples of this (like Lewis's 'Princeton = Princeton Borough') that would be mere foolishness. Evans's target is rather the view that there can be identity statements which are indeterminate in truth-value not because of any semantic indeterminacy but rather because of indeterminacy *in the world,* that is, *for no other reason* than that one or both of the objects *determinately* denoted by the singular terms flanking the identity sign is a vague object.

The natural conception of indefinite identity, so understood, and, I think, the conception which is Evans's target, can be elaborated as follows. An identity statement ‘$a=b$’ is true just in case every predicate true of a is true of b. It is false (and its negation is true) just in case there is some predicate true of a whose negation is true of b. If it is indeterminate whether $a=b$, then, where this is not due to semantic indeterminacy in one or both of the *singular terms* ‘a’ and ‘b’, it will not be the case that there is any predicate true of a whose negation is true of b, but neither will it be the case that every predicate true of a is also true of b, since there will be a predicate true of a which is neither true nor false of b.

What Evans's argument, as I understand it, demonstrates, is that this natural conception of indefinite identity can have no application. If ‘$a=b$’ is not true, whether it is false or merely indeterminate in truth-value, there *will* be a predicate true of a whose negation is true of b.

Evans employs the two operators ‘∇’ and ‘Δ’ to express indeterminacy and determinacy respectively. As already noted there is some unclarity about how he intends the latter operator to be understood, but the following interpretations will suffice for what follows: ‘∇p’ is to be true iff ‘p’ is neither true nor false, false iff ‘p’ is either true or false, in no case indeterminate in truth-value, ‘Δp’ is to be true iff ‘p’ is either true or false, false iff ‘p’ is neither true nor false, in no case indeterminate in truth-value (so understood ‘∇p’ should be read ‘it is indeterminate whether p’ and ‘Δp’ should be read ‘it is determinate whether p’).

Now suppose ‘$a=b$’ is neither true nor false, not because of any semantic indeterminacy in ‘a’ or ‘b’ but, in accordance with the conception of indefinite identity which is Evans's target, because of indeterminacy in the world. Then we have:

(1) $\nabla(a=b)$.

Since neither ‘a’ nor ‘b’ is semantically indeterminate we can move from this to:

(2) $\hat{x}[\nabla(x=a)]\,b$,

which ascribes to b the property expressed by ‘$\hat{x}[\nabla(x=a)]$’. But ‘$a=a$’ is determinately true, so we have:

(3) $\neg\nabla(a=a)$,

and from this, given the assumption that ‘a’ is determinate in denotation we can move to:

(4) $\neg\hat{x}[\nabla(x=a)]\,a$,

which ascribes to a the property expressed by ‘$\neg x[\nabla(x=a)]$’.

But given the assumption of the natural conception of indefinite identity outlined above that the existence of a predicate true of a whose negation is true of b *suffices* for the truth of '$\neg\, a = b$' we can now move from the truth of (2) and (4) to:

(5) $\neg\, a = b$,

contradicting the assumption, from which we began, that '$a = b$' is indeterminate in truth-value.

Moreover, given the explanation of 'Δ' above it is unproblematic to move from this to:

(5′) $\Delta \neg (a = b)$,

which is straightforwardly inconsistent with (1).

The above reasoning, I submit, provides the most powerful possible argument for Evans's conclusion; if it does not suffice to establish his conclusion nothing could.

How, then, might this reasoning be challenged?

Brian Garrett, in his [2], asserts that in Evans's argument his (1) and (5) are consistent and David Over, in his [6], appears to endorse this claim. But no such objection can be made to the argument just presented since, according to the conception of indefinite identity specified above as the target of the argument, the truth of '$\neg (a = b)$' *requires* the existence of a predicate true of a whose negation is true of b, whilst the truth of '$\nabla (a = b)$' *rules out* the existence of such a predicate.

Over also suggests that Evans's move from his (5) to his (5′) is problematic, but again, given the interpretation of the determinacy operator explained, no such problem attends the transition from (5) to (5′) in the argument above.

Another suggestion of Over's is that 'a believer in vague objects might believe that they cannot be rigidly or precisely designated', that is, that singular terms determinately denoting vague objects are impossible. If so, the believer in vague objects can say, he can block the moves from (1) to (2) and (3) to (4) in the argument above in just the way available to the proponent of the view that all indeterminacy in truth value is due to semantic indeterminacy, and in particular, that all indeterminacy in truth-value in identity statements is due to indeterminacy of denotation on the part of one or both of the singular terms flanking the sign of identity. Hence, he can conclude, the argument above fails to demonstrate any necessary incoherence in the idea of the existence of vague objects.

This point is a valid one, but it does not speak to the question at issue. For this is not whether the existence of vague objects is a coherent notion, but whether there can be identity statements which are indeterminate in truth value *for no other reason* than that one or both of the singular terms in them is a term *determinately* denoting a *vague object*, that is, whether there can be identity statements of indeterminate truth-value whose indeterminacy is *solely* due to an indeterminacy in the world, and not at all to any indeterminacy in language. If, whenever there is denotation of vague objects there is also *semantic* indeterminacy — indeterminacy in denotation — then the resultant indeterminacy in truth value, if any, can be explained in the way in which Lewis explains the indeterminacy of 'Princeton = Princeton Borough', and the idea that the objects indeterminately denoted are themselves *ontologically* indeterminate ceases to do any work in explaining the indeterminacy in truth-value. Whether it can ever do any work is then an open question. (It should be noted then that Garrett is wrong when he writes in his [2] 'The thesis that there can be vague objects is the thesis that there can be identity statements which are indeterminate in truth-value as a result of vagueness *the singular terms of which do not have their reference fixed by vague descriptive means*', for the existence of vague objects could be at most a necessary condition of the existence of identity statements which are indeterminate in truth-value in the way he characterizes.)

I believe that there is only one further objection that might be brought against the argument presented above, namely that the assumption that '$\neg a = b$' is *true* if there is a predicate true of a whose negation is true of b — which is the assumption which licenses the move from (2) and (4) to (5) — is unwarranted. What this assumption depends on is the principle which I wish to call the principle of the Diversity of the Definitely Dissimilar, which may be formulated as follows:

$$\text{(DDD)}\quad (\forall x)(\forall y)(\Delta Fx \,\&\, \Delta Fy \rightarrow (Fx \,\&\, \neg Fy \rightarrow \neg (x = y))).$$

It might seem that (DDD) is incontestable, but one philosopher who would certainly challenge this is Bruce Johnsen who formulates and objects to its equivalent:

$$\text{(LLnv)}\quad (\forall y)(\forall z)([\Delta \hat{x}(\phi x)]y \,\&\, [\Delta \hat{x}(\phi x)]z \rightarrow [(y = z) \rightarrow$$
$$(\hat{x}(\phi x)y \equiv \hat{x}(\phi x)z)])$$

in his paper [3].

But how can (DDD) (or (LLnv)) be regarded as objectionable? For all it says is that it is a sufficient condition of the distinctness of objects a and b that they be *definitely dissimilar*, and what can be the objection to that? If there is a predicate definitely true of a whose negation is definitely true of b how can it be indeterminate whether a is b?

But here, I think, we can see a last remaining bolthole for the proponent of indefinite identity. For he may say that indefinite identity is a kind of *relative* identity — a relation which ensures the indiscernibility of its terms in some, but not in all, respects, and in particular, not in respects expressible only by predicates containing '∇', or synonyms of such predicates. To demonstrate the incoherence of this notion it would be necessary to demonstrate that whenever objects are definitely dissimilar at all they must be definitely dissimilar in ways expressible without the use of such predicates. But neither Evans's original argument, nor the slightly more detailed version above, demonstrates any such thing.

This point is correct, but it remains, I submit, that the argument outlined above does refute the conception of indefinite identity I elaborated at the outset of the paper, which is, I suggest, not only a far more natural conception of indefinite identity than that of indefinite-identity-as-relative-identity but also the conception which was the target of Evans's original argument — indefinite-identity-as-relative-identity surely never entered Evans's mind.

I wish to make two final points about indefinite-identity-as-relative identity and the rejection of the principle of the Diversity of the Definitely Dissimilar to which a proponent of it is committed.

First, even though this notion cannot be shown to be definitely incoherent, it can be shown, I think (for the argument for this see my [5]), that the interpretation of any of the standard examples of indeterminately true identity statements in terms of this notion leads to conclusions which are much more unpalatable than those consequent upon the interpretation of such examples in terms of the notion of semantic indeterminacy.

My second point about indefinite-identity-as-relative-identity is the following. I have used the expression 'relative identity', but, of course, all I mean by this is *similarity* in certain respects; thus, similarity in respect of colour, for example is a relative identity relation, so is similarity in respect of shape and so on. So also are the two relations: *similarity in all respects except colour* and *similarity in all respects except shape.* These latter two relations, however, are logically quite on a par with indefinite-identity-conceived-as-relative-identity, that is: *similarity in all respects except those expressible only by predicates containing* '∇' *or synonyms of such predicates.* Let us focus now on the former of these two relations and, for easy reference, let us label it 'mere non-chromatic indistinguishability' (similarly, if we call predicates containing '∇' or 'Δ' 'delta predicates', and the associated properties 'delta properties', indefinite identity may be labelled 'mere non-delta property indistinguishability'). The fact that objects are merely non-chromatically indistinguishable, if this is allowed to be a coherent idea, is in no way incompatible with their definite non-identity, and that this is so is evident from the fact that objects between which this relation obtains will be *definitely dissimilar* in respect of colour. Indeed, if objects *a* and *b* are indistinguishable in all other respects, but not in respect of colour, not only will this not provide any reason for *doubting* their non-identity, it will in fact provide the strongest possible reason for *affirming* it, since then there will be a predicate, namely a colour predicate, true of *a* but false of *b*, and there can be no better ground for affirming non-identity than this. But what holds for the relation of mere non-chromatic indistinguishability holds also for indefinite-identity-conceived-as-relative-identity, that is, mere indistinguishability in non-delta properties. It, too, as we have seen, can obtain only between objects which are *definitely dissimilar* in certain respects, namely in respect of their delta properties, so it too is such that its obtaining cannot be regarded as in any way incompatible with definite non-identity. Indeed, we have as good a ground for regarding as distinct objects which are merely non-delta property indistinguishable as we have for so regarding objects which are merely non-chromatically indistinguishable, namely the fact that there will be a predicate, viz. a delta predicate, true of one but false of the other. What all this suggests is that mere non-delta property indistinguishability is no more appropriately thought of as a kind, or even a borderline case, of *identity* (hence as 'indefinite identity') than is the logically analogous relation of mere non-chromatic indistinguishability. Like the latter relation it is simply a similarity relation, and its coherence or otherwise is as irrelevant to the question of vague *identity* as is the coherence or otherwise of the idea of mere non-chromatic indistinguishability.[1]

But whether or not this last point is found convincing, the point I wish to emphasize, to conclude, is that the defence of the coherence of the notion of indefinite-identity-as-relative-identity and its relevance to the topic of vague identity, and the consequent rejection of the principle of the Diversity of the Definitely Dissimilar, is the *only* road forward for an opponent of Evans's position.

University of Birmingham,
P.O. Box 363, Birmingham B15 2TT

---

[1] In fact, of course, in either case the question about coherence is a question about supervenience: do colour properties (or: delta properties) supervene on other properties or can they vary independently?

REFERENCES

[1] Gareth Evans, 'Can There Be Vague Objects?', *Analysis* 38 (1978) 208
[2] B. J. Garrett, 'Vagueness and Identity', *Analysis* 48 (1988) 130–35
[3] B. Johnsen, 'Is Vague Identity Incoherent?', *Analysis* 49 (1989) 103–12
[4] D. Lewis, 'Vague Identity: Evans Misunderstood', *Analysis* 48 (1988) 128–30
[5] H. W. Noonan, 'Indefinite Identity: A Reply to Broome', *Analysis* 44 (1984) 117–21
[6] D. E. Over, 'Vague Objects and Identity', *Analysis* 49 (1989) 197–99

CONTINGENT IDENTITY IN COUNTERPART THEORY

By MURALI RAMACHANDRAN

I WISH to propose a slight modification to the translation scheme for counterpart theory I put forward in [3]. The motivation for this change is that it makes for a more plausible account of contingent identity.

I

First, let us look at David Lewis's original theory of counterparts (CT), advanced in [1], for comparisons. The formula

(1) $(x = y) \mathbin{\&} \Diamond(x \neq y)$

— presumably read simply as 'x is contingently identical with y' — comes out true on a model if and only if $\mathrm{Ref}(x) = \mathrm{Ref}(y)$ and $\mathrm{Ref}(x)$ has two or more counterparts in some world. Thus, contingent identity here amounts, metalogically speaking, to contingent *uniqueness*: an object **a**, is said to be contingently identical with something just in case **a** 'might have been' two or more objects.[1]

This results in a fairly toothless account of contingent identity — witness the CT-validity of

(K) $(x = y) \supset (\Diamond \alpha \equiv \Diamond \alpha[y/x])$

and

(L) $\Box(x = y) \supset ((x = z) \supset \Box(x = z))$

which I discussed in [3], pp. 138–9.[2] (K) entails that contingently identical objects must have the same 'possible' properties, so that whatever property one object might have had, the other might have had as well. And (L) entails that if an object, **a**, is necessarily identical with something, then nothing is contingently identical with **a**. Also,

(2) $(x = y) \supset \Box(\exists z(z = x) \equiv \exists z(z = y))$

comes out valid, so that for any contingently identical objects, one exists in a world iff the other exists there as well. This seems an implausible requirement.

---

[1] 'Metalogically speaking' because, strictly speaking — on the rule that 'a might have been F' is true iff Ref(a) has a counterpart that is F — something might have been two objects only if one of its counterparts is two objects, which is impossible.

[2] Graeme Forbes has pointed out an error of mine regarding (K). I claimed that (K) was CT-valid 'where "$\alpha[y/x]$" is the result of replacing every free occurrence of "x" in α with "y" — as long as "y" does not become bound in the process' ([3], p. 138). But (K) is false if α is the formula '$\Box(x \neq y)$', because '$\Box(y = y)$' is CT-valid. So let me claim instead that (K) is CT-valid where 'y' does not occur in α. My misgivings about (K) are still pertinent.

One result, though, seems quite plausible:

(3) $\Box(x=x)$

is CT-valid, and entails that no object is contingently self-identical. Like Lewis, I find the notion of contingent self-identity incoherent; but (3) is not valid in the system I presented, CT*. The modification I am going to propose will render (3) valid. In any case, it should be pointed out that even though (3) is CT-valid,

(4) $\exists x(\Box(x=x) \,\&\, \forall y((x=y) \supset \sim\Box(x=y)))$

comes out CT-satisfiable. (4) reads: there is an object which is necessarily identical with itself and (yet) no object identical with it is necessarily identical with it![3] I don't know how to make sense of this possibility.

So, all things considered, I contend that CT accommodates an unconvincing account of contingent identity.[4]

II

In [3], *atomic constituents* of sentences were generally translated at the 'atomic' translation stage according to one rule (see pp. 134–5). I now propose preferential treatment for those atomic constituents of the form 'a = b': we replace any atomic constituent of the form 'a = b' — even where 'a' and 'b' are the same term — at the atomic translation stage with '$\forall x(Cxa \equiv Cxb)$'. Thus, intuitively, 'a = b' is true in a world just in case Ref(a) and Ref(b) have precisely the same counterparts in that world (even if neither has a counterpart in that world);[5] and '$\Box(a=b)$' comes out true just in case Ref(a) and Ref(b) have the same counterparts, period.

That's it: modification over. (3) comes out CT*-valid now because Ref(x) must have the same counterparts as Ref(x) in every possible world. (3) dictates that (4) is equivalent to

(5) $\exists x\forall y((x=y) \supset \sim\Box(x=y))$,

which is incompatible with a CT*-valid formula,

(6) $\forall x\forall y((x=y) \supset \Box(x=y))$.

---

[3] The cause of this perplexing result is that each variable-*type*, rather than each variable-*token*, occurring free within the scope of a modal operator introduces a counterpart-quantifier in translation.

[4] Of course, Lewis has revised his original counterpart theory so that it now incorporates a multiplicity of counterpart relations (see, e.g., [2]). An exploration of the accompanying account of contingent identity must be left to another paper.

[5] Identity really amounts to *indistinguishability* here. If two objects have precisely the same counterparts in a world then, in an important sense, they are indistinguishable in that world. For, whatever (non-modal) property one has at that world, the other has there as well.

((6) is valid because objects from any one world which are identical must have — in Lewis's counterpart theory as well as in mine — precisely the same counterparts in all worlds.) Hence, (4) is CT*-unsatisfiable, as desired.

Now, the validity of (6) dictates that no objects from any one world can be contingently identical. Moreover, (1) cannot be regarded as the correct symbolization of 'x is contingently identical with y' in CT*, for '$x=y$' comes out true even when Ref(x) and Ref(y) have no counterpart in the actual world. Indeed, (1) comes out true whenever Ref(x) and Ref(y) are any two objects from a possible world that have no *actual* counterparts.[6] So it might well be wondered how contingent identity can raise its ugly head in CT* at all.

Fortunately, the special existence-predicate, 'E', introduced on p. 136 of [3], enables us to ensure the existence of actual counterparts. The content of sentences of the form 'a is contingently identical with b' is captured by formulae of the form:

(7) $Ea \& Eb \& (a=b) \& \Diamond(a\neq b)$.

Likewise, for the record, sentences of the form 'a is contingently non-identical with b' are symbolically expressed as

(8) $(a\neq b) \& \Diamond(Ea \& Eb \& (a=b))$.

'$\Diamond(a\neq b)$', and therefore (7), can be satisfied only if Ref(a) and Ref(b) are either possible objects from different worlds or distinct objects from a strictly possible (i.e., unactual) world. Thus, if (7) is to be true, Ref(a) and Ref(b) *have to be* distinct possible objects.

So on this account of contingent identity, we can coherently maintain the contingent identity of *distinct* possible objects, but cannot coherently maintain the contingent identity of an actual (or possible) object and *that same* object. Contingent self-identity is therefore ruled out, as desired.

III

Now, I should point out two *prima facie* anomalies. First, something similar to (4), namely (9):

(9) $\Box(x=x) \& \forall y((x=y)\supset \sim\Box(x=y))$

— read 'x is necessarily identical with itself and (yet) nothing identical with it is necessarily identical with it' — turns out to be satisfiable in CT*. ((9) is clearly CT-satisfiable, since (4) is.) But (9), unlike (4), is bearable once we realize that it is satisfied in CT* only if Ref(x) is an unactual object and the universal-quantifier quantifies only over actual objects. Indeed, (9) is equivalent in CT* to (10):

---

[6] (1) would not have been adequate in the original CT* either, since '$\Diamond(x\neq y)$' is satisfied even when Ref(x) or Ref(y) have no counterparts in some world. Indeed, (1) comes out true in the original CT* whenever Ref(x) is Ref(y) and Ref(x) has no counterpart in some world.

(10) $\Box(x=x)\ \&\ @\forall y((x=y) \supset \sim\Box(x=y))$

— x is necessarily identical with itself and no *actual* object identical with it *in the actual world* is necessarily identical with it — which does not seem incoherent.

The second *prima facie* anomaly is that (11):

(11) $\Box(x=y)$

— x is necessarily identical with y — can be satisfied even though Ref(x) and Ref(y) are distinct possible objects! Here's the mitigatory reason. A sentence is necessarily true just in case it is true in every possible world; objects from different worlds, **a** and **b**, say, even though they are distinct possible objects, may nevertheless have the same counterparts in every world; thus '$\underline{\mathbf{a}}=\underline{\mathbf{b}}$' (where '$\underline{\mathbf{a}}$' and '$\underline{\mathbf{b}}$' are names of **a** and **b** respectively) may be true in every world, i.e., necessarily true. The anomaly dissolves because we can make sense of the claim that there is no world *with respect to which* **a** is not identical with **b**, even though **a** and **b** are distinct possible objects. Note that (11) is satisfied in CT too under such circumstances.

Regarding the drawbacks of CT mentioned on p. 163, it can be checked that, on the present account of contingent identity: contingently identical objects can differ with respect to 'possible' properties; an object can be necessarily identical with one thing but only contingently identical with another; and contingently identical objects needn't inhabit precisely the same worlds.

Finally, note that my modification to CT* alters only one of the results in the 'validity-table' on p. 138 of [3]. (J) (=(3) in this paper) is now CT*-valid; but to my mind, as already confessed, this result is to be welcomed.

Birkbeck College,
Malet Street, London WC1E 7HX

REFERENCES

[1] David Lewis, 'Counterpart Theory in Quantified Modal Logic', in his *Philosophical Papers*, Vol. 1 (Oxford: Oxford University Press, 1983), 26–46.

[2] David Lewis, *On the Plurality of Worlds* (Oxford: Basil Blackwell, 1986).

[3] Murali Ramachandran, 'An Alternative Translation Scheme for Counterpart Theory', *Analysis* 49 (1989) 131–41.

COUNTERPARTS, LOGIC AND METAPHYSICS: REPLY TO RAMACHANDRAN

By GRAEME FORBES

IN a recent paper in ANALYSIS [10], Murali Ramachandran presents an ingenious version of counterpart theory which stays much closer to Lewis's original approach [7] than did my variant of it [1], but which at the same time solves the problem of expressing strong necessity (inability to express strong necessity was the main drawback to Lewis's system). In addition, Ramachandran uncovers some problems with my so-called 'canonical' or 'free' counterpart theory (CCT) in the light of which he argues that his modification (CT*) of Lewis's system CT is superior to CCT. In this note, I will argue for two claims. My first claim is that, even accepting Ramachandran's main objection to CCT, the difficulties can be circumvented by a technical fix. My second claim is that Ramachandran's treatments of identity, predication and Leibniz's Law in [10] are inadequate, and remain so despite his subsequent modification of CT* [11]. In addition, in pursuing this latter point I will suggest a reason to doubt that my original version of CCT misrepresents the logic of '$\Box$', despite Ramachandran's examples.

I A REVISION TO CCT

Ramachandran points out that when names of merely possible objects are allowed into the modal language, the schema $\ulcorner\Box A\supset A\urcorner$ has counterinstances in CCT, and if the modal language contains the actuality operator @, then $[@A\equiv A]$ has counterinstances (for these purposes, a model verifies a counterinstance to a schema if the instance holds at the actual world w^* of the model).[1] The basic stipulation governing the counterpart relation in CCT is:

(C) For any object x, x is a counterpart of x at w iff either x exists at w or x has no existing counterparts at w; and if x is a counterpart of x at w, nothing else is.

Suppose 'β' is a name of an object which does not actually exist. Suppose also that (a) at every world, β has counterparts which exist and all of which satisfy the arbitrarily complex but non-modal condition ϕ; and (b) β itself does not exist at, nor satisfy ϕ at, the actual world w^*. Then by (a), $\ulcorner\Box\phi\beta\urcorner$ holds at the actual world, since at every world u, every counterpart of β at u satisfies ϕ at u (by (C), fact (b) does not prevent the truth of $\ulcorner\Box\phi\beta\urcorner$ at w^*).

---

[1] In Hodes' terminology in [5], $\ulcorner@A\equiv A\urcorner$ is valid but not strongly valid. By the system S5@ I will mean the axiomatization of the valid formulae which Hodes gives in [5].

On the other hand, by (b), $\ulcorner\phi\beta\urcorner$ is false at w^*. This is because a non-modal sentence holds at a world according to CCT iff it holds at that world according to standard Kripke semantics: the counterpart relation is not appealed to. So we have a counterinstance to $\ulcorner\Box A \supset A\urcorner$. The counterinstance to $\ulcorner @A \equiv A\urcorner$ is essentially the same ([10], p. 139).

The problem arises because the counterpart relation is not appealed to at a point where it should be. We can dissolve the problem, therefore, simply by insisting that the counterpart relation be appealed to in the appropriate way. Translating a sentence in CCT eventually reduces to translating atomic formulae of the general form $\ulcorner Ft_1 \ldots t_n\urcorner$. Suppose that in the last stage of translation, when atomic formulae are being relativized to a world-term ζ, for each term-occurrence t_i in $\ulcorner F_1 \ldots t_n\urcorner$ we write $\ulcorner \exists v_i C v_i t_i \zeta \ \&\urcorner$ and substitute v_i for t_i in $\ulcorner Ft_1 \ldots t_n\urcorner$. For the formula $\ulcorner \Box F\beta \supset F\beta\urcorner$, using the clause for $\Box$ in [1], we would then obtain

$$(1)\quad [\forall w \forall x (Cx\beta w \supset \exists y (Cyxw \ \& \ Fyx))] \supset \exists x (Cx\beta w^* \ \& \ Fxw^*)$$

Since CCT guarantees at least one counterpart of β at any world w, and also guarantees that that counterpart is its own sole counterpart at w, this conditional is a theorem of CCT, since any y as in (1) which is a counterpart of x at w must *be* x.

This modification by itself is insufficient to preserve $\ulcorner @A \equiv A\urcorner$ if we retain clause (17) for @ from my [2]. But in the context of our new treatment of atomic formulae, the difficulty is circumvented if we drop that clause, so that the only effect of @ is to change the world-term of relativization. The essentials of the revised translation scheme for CCT which I propose, therefore, are these (see [1] for (fussier) notation, restrictions etc.):

(a) If $Ft_1 \ldots t_n$ is atomic, $\mathrm{Rel}[(Ft_1 \ldots t_n), w] =$
$\exists v_1 \ldots \exists v_n (Cv_1 t_1 w \ \& \ldots \& \ Cv_n t_n w \ \& \ Fv_1 \ldots v_n w)$
where $v_1 \ldots v_n$ do not occur in $Ft_1 \ldots t_n$;
(b) $\mathrm{Rel}[\exists vA, w] = \exists v(Evw \ \& \ \mathrm{Rel}[A, w])$;
(c) $\mathrm{Rel}[\Box A, w] = \forall w' \forall v_1 \ldots \forall v_n (Cv_1 t_1 w' \ \& \ldots \& \ Cv_n t_n w' \supset$
$\mathrm{Rel}[A(t_i/v_i), w']$
if $t_1 \ldots t_n$ are the occurrences of constants or free variables in A not within the scope of any modal operator ($\Box$, $\Diamond$. @) in A, $\mathrm{Rel}[\Box A, w] = \forall w' \mathrm{Rel}[A, w']$ if there are no such terms;
(d) $\mathrm{Rel}[@A, w] = \mathrm{Rel}[A, w^*]$;
(e) For every sentence A, $\mathrm{Trans}[A] = \mathrm{Rel}[A, w^*]$.

It is clear that we now have $\ulcorner @A \equiv A\urcorner$, since by (d) and (e) the first steps in translating the two sides of the biconditional reduce each side to $\mathrm{Rel}[A, w^*]$. In fact, clauses (d) and (a) make @ behave in a much more orthodox way than did my treatment of it in [2]. Thus it is no longer the case that

$$(2)\quad \Diamond \exists x @ \exists y (x = y \ \& \ @Fx \ \& \ @{\sim}Fx)$$

is satisfiable in CCT, since the translation of (2) is now

$$(3)\quad \exists w\exists x(\exists xw \,\&\, \exists y(Eyw^* \,\&\, \exists x'\exists y'(Cx'xw^* \,\&\, Cy'yw^* \,\&\, x'=y' \\ \&\, \exists x'(Cx'xw^* \,\&\, Fx'w^*) \,\&\, {\sim}\exists x'(Cx'xw^* \,\&\, Fx'w^*)))).$$

The last two conjuncts are mutually contradictory, so (3) is unsatisfiable and therefore (2) is logically false. More generally, @ conforms to Hodes' axioms (@1), (@2) and (@3) for actuality ([5], p. 28). Furthermore, using the obvious equivalence between models of S5@ and models of CCT in which each object has exactly one counterpart at any world and no two objects share a counterpart at any world – that is, models in which the counterpart relation is *identity-like* – it follows that CCT is conservative over S5@: no non-theorem of S5@ is translated into a theorem of CCT. And if we add to CCT the stipulation that the counterpart relation be identity-like, then every theorem of S5@ is translated into a theorem of CCT. So in the general case, where there are validities of S5@ which are not translated into theorems of CCT, this is entirely due to allowing the counterpart relation to be less than identity-like. So far as I am aware, CCT is the only counterpart theory which meets the adequacy condition of coinciding with Kripke semantics for a modal language without @ when the counterpart relation is identity-like.[2]

II Predication

One area of disagreement between CT* and CCT concerns the Falsehood Principle, which in Kripke semantics states that an n-place atomic predicate is false of an n-tuple of objects at a world w if any element of the tuple does not exist at w. In the standard Kripke semantics for S5 the Falsehood Principle is not imposed, so that no instance of e.g. $\ulcorner(\forall x)\Box(Fx\supset Ex)\urcorner$ for atomic F is a theorem. But it is a theorem of CT and CT*. In [3] I argued that this is a defect in CT, on the grounds that whether or not the Falsehood Principle is correct is a metaphysical question which logic should not foreclose. To this Ramachandran replies that *any* theorem forecloses something ([10], p. 140). But the question, of course, is whether it is a logical, or merely a metaphysical, falsehood which is ruled out. And the issues surrounding the Falsehood Principle seem clearly metaphysical, reaching far beyond a discussion of what the inferential behaviour of □ and ◇ is; the Falsehood Principle may even be *false* (see [4], Ch. 3). At any rate, we do reject modal logics on just these kinds of ground. Perhaps the following is false: it is possible for immaterial entities to exist, but impossible for any actual entity to be immaterial. Nevertheless, it is not logically false, and a semantics which validates $\ulcorner\Diamond\exists xFx\supset\exists x\Diamond Fx\urcorner$ is thereby unacceptable. The Falsehood Principle is on an equal footing. Indeed, Ramachandran now concedes that the Principle is false for the logical atomic relation symbol '$=$', since the effect of his modification of CT* in [11] is to allow '$x=y$' to be satisfied at w by an object with no counterpart amongst the existents of w. But what holds for identity may well hold for other abstract properties and relations.

---

[2] The original version of CCT satisfies this condition so long as names of non-actual objects are excluded from the language.

III Contingent Identity and Leibniz's Law

In Ramachandran's original CT*, presented in [10], and also in my CCT, the simplest form of the necessity of identity, which in CCT is

(4) $\alpha=\beta\supset\Box(\alpha=\beta)$

is invalid. In both systems, $\ulcorner(\forall x)\Diamond(x\neq x)\urcorner$ is satisfiable, which is why the standard S5 proof of (4) is unsound, since that proof has $\ulcorner(\forall x)\Box(x=x)\urcorner$ as a premiss ([6], p. 136). But this does not mean that the counterpart theorist is committed to the idea that an object α could have been non-self-identical (written $\ulcorner\Diamond\lambda x(x\neq x)[\alpha]\urcorner$). The possibility that $\alpha\neq\alpha$, written $\ulcorner\Diamond(\alpha\neq\alpha)\urcorner$ or $\ulcorner\Diamond\lambda x\lambda y(x\neq y)[\alpha,\alpha]\urcorner$, is rather the possibility for α of being two or more things. For example, it might be said that if the parts which make up this clock had gone half into one clock and half into another, then this clock would have been two clocks; in modal language, the thesis is

(5) $\exists x\Diamond\exists y\exists z(x=y\ \&\ x=z\ \&\ y\neq z)$

and so $\ulcorner\exists x\Diamond(x\neq x)\urcorner$.[3] But in Ramachandran's CT*, (5) is logically false: $\ulcorner x=y\urcorner$ implies every counterpart of x is identical to y, while $\ulcorner x=z\urcorner$ implies every counterpart of x identical to z, so $\ulcorner y\neq z\urcorner$ would hold in CT* only if x had no counterparts. It is hard to see how to square this with a willingness to allow the logical possibility of contingent identity.

Perhaps Ramachandran has some other conception of contingent identity which does not demand the logical possibility of (5). I consider CT* in both the original and revised version. The example he gives in [10] is that his mind and his brain are contingently identical, his brain occupies his body rigidly, but his mind does not. I take this to mean that there is a single entity, variously referred to as Ramachandran's mind or Ramachandran's brain, and a possibility for that entity of being two things, one housed in a certain body, the other not. But this does not mean that there are possibilities for Ramachandran's (actual) mind which are not possibilities for his (actual) brain: since that mind *is* that brain, the counterpart arrows which lead from it in the actual world to a mind, and also to a brain in a body not housing that mind, in a single non-actual world w, lead *equally* from the actual mind and the actual brain. So if that mind has the possibility of not being housed in a certain body, there is the same possibility for that brain (this argument is just substitution of identicals in the counterpart-theoretic translations).[4] Perhaps the irrelevant *de dicto* readings of the example confused the issue in [10].

---

[3] Perhaps 'contingent unity' is a better label for the phenomenon than 'contingent identity'. But I see no objection to rendering 'this thing could have been two things' as 'there could have been two things such that this thing is identical to one and to the other'. In my view this could only be the case to some intermediate degree; see ([3], Ch. 7).

Apparently this is not how Ramachandran saw matters in [10], since there he regards the mind/brain case as a refutation of this instance of Leibniz's Law:

$$(7) \quad \alpha = \beta \supset [\Diamond \phi(\alpha) \equiv \Diamond \phi(\beta)]$$

which he says is invalid in CT*. Unfortunately, he does not supply an invalid instance, and there are certainly none corresponding to the mind/brain example if the analogue of (C) for CT* is in force. If 'α' and 'β' are names of the same worldbound individual — whether actual or not — then $\ulcorner \Diamond \phi(\beta) \urcorner$ translates into a sentence which can be inferred from the translation of $\ulcorner \Diamond \phi(\alpha) \urcorner$ by Identity Elimination, since Ramachandran's translation scheme, like that of CCT, introduces a counterpart qualifier for each occurrence of a term, and substitution preserves the number of occurrences of terms, though not the number of term-types which have occurrences. And since the language of counterpart theory is a classical first-order language, this application of IE is unobjectionable.[5]

For all that, (7) *does* have counterinstances in CT* (both versions) and in CCT as revised in §1 above. (7) can fail when 'α' and 'β' are names of distinct non-actual worldbound individuals which have one and the same counterpart in the actual world. (This does not save Ramachandran's mind/brain case, where the rigid designators stand for the same actual object.) And in [11], Ramachandran modifies CT* so that (4) can fail only when 'α' and 'β' are such names and have the same counterparts at w*, though he does not explain what the *point* of the resulting notion of contingent identity is: some versions of the necessity of identity come out valid, others invalid, depending on exactly what the quantifier-operator prefix is. However, the general moral seems to be that familiar validities do not hold when names of non-actual objects are allowed into the language.

---

[4] Lewis would resist this argument by claiming that the sortals 'mind' and 'brain' invoke two different counterpart relations; see [8].

[5] Similar reasoning establishes the equivalence if the names are construed as names of transworld individuals ([9], 4.3) with identical actual-world parts.

This observation takes us back to Ramachandran's main objection to CCT, that it allows ⌜$\Box A \supset A$⌝ to have counterinstances. Since the counterinstances involve names of non-actuals, we could reasonably ask why the failure of ⌜$\Box A \supset A$⌝ in CCT when these are allowed is worse than the corresponding failure of (4) and (7) in Ramachandran's systems. However, I do not wish to leave matters with an *ad hominem* rhetorical question. Rather, it seems to me that there are philosophical difficulties with the sorts of names we are countenancing here. Our exemplar of the introduction of a name of a non-actual into a language is by a reference-fixing stipulation of the form 'let "a" denote the thing which would have resulted if process p had unfolded', where process p would determine an individual essence. If there are no individual essences, this stipulation will not work even in Kripke semantics. And it will not work in counterpart-theoretic semantics even if there are individual essences, since individuals are worldbound: *which* of the worldbound individuals which in different worlds represent the thing which would have resulted if process p had unfolded would be denoted by 'a'? This problem appears to persist no matter how the content of process p is expanded. Apparently, then, we can accommodate names of non-actuals only by positing essences and reconstruing the ontology of the modal language as one of transworld individuals, which, at *best*, is compatible only with Lewis's modal realism (see [9], 4.3).

If names of non-actuals are excluded from the modal language, the counterexample to ⌜$\Box A \supset A$⌝ which prompted the revisions to CCT in §1 lapses, while if we preserve the treatment of @ in §1 and clause (a) only for atomic formulae within the scope of an @ and no intermediate modal operator (using (1) of [1] otherwise), we obtain a correct logic of 'actually'. I believe this hybrid of the original and revised versions of CCT is optimal in matching the divergence of the theorem-set from the theorems of S5@ with the divergence of the counterpart relation from an identity-like relation.[6]

Tulane University,
New Orleans, LA 70118, U.S.A.

REFERENCES

[1] Graeme Forbes, 'Canonical Counterpart Theory', *Analysis* 42 (1982) 33–7.
[2] Graeme Forbes, *The Metaphysics of Modality* (Oxford: Oxford University Press, 1985).
[3] Graeme Forbes, 'Free and Classical Counterparts: Response to Lewis', *Analysis* 47 (1987) 147–52.
[4] Graeme Forbes, *Languages of Possibility* (Oxford: Basil Blackwell, 1989).
[5] Harold Hodes, 'Axioms for Actuality', *The Journal of Philosophical Logic* 13 (1984) 27–34.

[6] I thank Murali Ramachandran for comments on an earlier version of this paper.

[6] Saul Kripke, 'Identity and Necessity', in *Identity and Individuation* edited by Milton K. Munitz (New York: New York University Press, 1971), 135–64.

[7] David Lewis, 'Counterpart Theory and Quantified Modal Logic', in his *Philosophical Papers* Volume 1 (New York: Oxford University Press, 1983), 26–46.

[8] David Lewis, 'Counterparts of Persons and Their Bodies', in his *Philosophical Papers* Volume 1, 47–54.

[9] David Lewis, *On The Plurality of Worlds* (Oxford: Basil Blackwell, 1988).

[10] Murali Ramachandran, 'An Alternative Translation Scheme for Counterpart Theory', *Analysis* 49 (1989) 131–41.

[11] Murali Ramachandran, 'Contingent Identity in Counterpart Theory', *Analysis*, this issue.

UNSUCCESSFUL REVISIONS OF CCT

By MURALI RAMACHANDRAN

IN [5], I presented a counterpart-theoretic semantics for modal logic (CT*) that interpreted necessity 'strongly' without admitting certain implausible features of Graeme Forbes's canonical counterpart theory (CCT) (see [1] and [2]). Forbes has responded [3], his central aim being to show that the problems for CCT I mentioned in [5] — e.g., the satisfiability of (1) and the invalidity of (2):

(1) $\Diamond(\exists x)(@(\exists y)(x=y) \;\&\; (@Fx \;\&\; @\sim Fx))$

(2) $\Box\alpha \supset \alpha$

— can be 'circumvented by a technical fix' (p. 167). Indeed, he provides two ways of so fixing CCT. I argue here that these revisions do not avoid all the problems. In particular, we shall see, there is a problem for CCT, as 'bad' as (1)'s satisfiability, that neither revision, nor any obvious modification of these revisions, avoids. In so far as this problem seems endemic to canonical counterpart theory, I contend that CT* is the better alternative.

I

But first, a few words about Forbes's misgivings about CT*.

Leibniz's Law. In [5], I suggested that the CT*-invalidity of Leibniz's Law:

(3) $\alpha=\beta \supset [\Diamond\phi(\alpha) \equiv \Diamond\phi(\beta)]$

does not tell against CT*, because it is not implausible that *contingently* identical objects have different properties at some world (where they are not identical). As Forbes points out, my mind/brain example (p. 138) does not refute (3), since 'my mind' and 'my brain' are rigid designators of the same object. I concede the point (though, for the record, the example was intended to challenge only the *obviousness* of (3)). Nevertheless, seeing as

(4) $\Diamond\exists x\exists y(x \neq y \ \&\ Fx \ \&\ \Box{\sim}Fy \ \&\ @(x=y))$

is satisfiable in all versions of CCT — viz. when two objects from an unactual world are such that: one of them is F (and so, possibly F) while the other is necessarily non-F, and they are (contingently) identical in the actual world — I can't see that (3)'s invalidity in CT* can be deemed a telling point against it.

Contingent Identity. Forbes finds it hard to square the unsatisfiability of

(5) $\exists x\Diamond\exists y\exists z(x=y \ \&\ x=z \ \&\ y\neq z)$

in CT* with my willingness to allow the logical possibility of contingent identity. Certainly, as he points out, CT* cannot accommodate the claim that my mind and my brain are contingently identical (p. 171). However, as he acknowledges (note 3, p. 170), contingent identity on his account amounts to contingent 'unity' — (I used the phrase 'uniqueness' in [6]) — whereas I prefer to employ 'contingent identity' to signify a different phenomenon, a more likely candidate in my opinion, and one which doesn't require the satisfiability of (5) (see [6]).

Names of Non-Actual Objects. Forbes thinks there are philosophical difficulties about countenancing names of non-actual objects in our logic. Since possible objects are world-bound individuals in counterpart theory, any attempt to anchor a name to a non-actual object with a reference-fixing description will fail because *many* world-bound individuals will satisfy that description (if any does). I see a way out. If — like David Lewis [4] — we think there can be a *similarity-ordering* of possible worlds, so that some world can be regarded as being 'closer' (more similar) to the actual world than another, then this problem is easily rectified: we can anchor the name to the 'closest' object that satisfies the description (if there is one). And note, *contra* Forbes (p. 172), that this tack would not require our positing *essences.* But in any case, as we shall see, the crucial differences between CCT and CT* are independent of this issue: even excluding names of non-actuals does not get Forbes's counterpart theory out of the woods.

The Falsehood Principle. Forbes takes exception to CT*'s imposition of the Falsehood Principle; his view is that it is a *metaphysical* principle and that logic should not foreclose metaphysical (as opposed to logical) questions. I am not so sure what kinds of claims logic should or should not foreclose. After all, any serious modal logic (including CCT) seems inevitably to foreclose metaphysical questions. Indeed, Forbes revises CCT because it *doesn't* foreclose the question of (1) and (2)'s truth. Of course, Forbes's underlying point here may be that it is far from clear that the principle is true, and that it therefore has to be supported. Support will not be forthcoming in this paper. I leave it to the reader to choose between the drawbacks of the principle and the drawbacks of CCT and its cousins.[1]

II

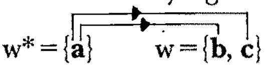
As already mentioned, Forbes presents two ways of revising CCT to avoid its problems. The first way is outlined on p. 168 (clauses (a)–(d)). The resulting system, call it *Modified*-CCT (MCCT), differs from CCT with respect to its treatment of *atomic* formulae (viz., clause (a)) and sentences governed by the actuality operator (viz., clause (d)). Later, because of the alleged difficulties surrounding names of non-actual objects, Forbes proposes the exclusion of such names. Then, he suggests, his original theory, CCT, with clause (d) of MCCT's scheme – and sometimes clause (a) (p. 172) – also avoids the problems. This hybrid of CCT and Modified-CCT, call it *Hybrid*-CCT (HCCT), seems to be Forbes's preferred alternative to CCT (p. 172). But, as we will see, both revisions fail.

Starting with HCCT, there is one problem for CCT I mentioned in [5] which Forbes seems to have overlooked. I suggested there (p. 132) that the satisfiability of

(6) $\Diamond(Fx \mathbin{\&} {\sim}Fx)$.

– read: it might have turned out both that x is F and that x is not F – in CCT was at least as unpalatable as the satisfiability of (1). Although HCCT circumvents the problems posed by (1) and (2), (6) remains satisfiable. Here's a satisfying model:

(M1)



$$\mathrm{Ref}(x) = \mathbf{a}$$
$$\mathrm{Ext}(F, w^*) = \{\langle \mathbf{a} \rangle\}$$
$$\mathrm{Ext}(F, w) = \{\langle \mathbf{b} \rangle\}$$

(The arrows trace the counterpart relation: so **b** and **c** are counterparts of **a** at w. Forbes's restriction on names is not violated since Ref(x) is an actual object.) Seeing as (1)'s satisfiability in CCT (partially) motivated Forbes's revision of CCT, (6)'s satisfiability in

---

[1] Forbes takes me to be abandoning the principle in [6], where I modified CT* to render '$\Box(x=x)$' valid. For, on the modified scheme (MCT*), '$x=y$' is satisfied at a world when neither Ref(x) nor Ref(y) exists (has a counterpart) in that world. I must confess that I did not have the Falsehood Principle in mind when making my modification (and so did not intend to abandon it). However, MCT*'s treatment of identity does not *commit* one to rejecting the principle. For we can take *distinctness* (i.e. non-identity), rather than identity, as the primitive relation and define identity as the negation of non-identity. Fortunately, in MCT* '$x \neq y$' can't be satisfied at a world unless Ref(x) and Ref(y) exist in that world.

HCCT surely provides as much motivation for revising (if not rejecting) HCCT.[2]

HCCT also embraces an unprincipled metaphysical difference between the actual world (or objects) and unactual ones. It is permissible for *actual* objects to be F and not be F in some (unactual) world:

(6*) $\exists x \lozenge (Fx \;\&\; {\sim}Fx)$

being satisfiable — but not permissible for *unactual* objects to be F and not be F in the actual world:

(7) $\lozenge \exists x @ (Fx \;\&\; {\sim}Fx)$

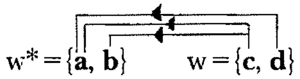
being unsatisfiable. Surely *this* is a metaphysical question which logic shouldn't foreclose!

MCCT avoids the problems mentioned so far but renders

(8) $((x = y) \;\&\; (y = z)) \supset (x = z)$

— the transitivity of identity — invalid. Here's a counterexample model:

(M2)



Ref(x) = **b**; Ref(y) = **c**; Ref(z) = **d**

We only need clause (a) of MCCT to evaluate (8). '$x = y$' is true because Ref(x) and Ref(y) both have **b** as a counterpart in w*; '$y = z$' is true because Ref(y) and Ref(z) both have **a** as a counterpart in w*; but '$x = z$' is false because Ref(x) and Ref(z) have no common counterpart in w*. MCCT may boast to matching Kripke-semantics when the counterpart relation is identity-like (p. 169), but at least in CT* *identity* is identity-like on all occasions!

One obvious way of securing the validity of (8) is to simply impose HCCT's restriction on names, that is, to preclude names of non-actual objects in MCCT. Call the resulting system *Restricted-*MCCT (RMCCT). However, RMCCT is vulnerable to the following problem, a problem that CCT can't seem to shake off.

(9) $\lozenge (x = x \;\&\; x \neq x)$.

— it might have turned out both that x was identical with x (i.e. itself?) and that it wasn't — is satisfiable in all the aforementioned versions of CCT. (M1) above suffices as a satisfying model for all cases.[3] (9) is surely unacceptable: it is no better than (1) and (6). Forbes's suggestion that we read

---

[2] This also foils another of Forbes's aims in [3]: to give a reason to doubt that his original version of CCT misrepresents the logic of '□' (p. 167). His idea was that counter-intuitive results about necessity (possibility) would not arise once we exclude names of non-actuals, since this tack does avoid (2). Unfortunately, as (6) shows, CCT has deeper problems.

(10) $\Diamond(\alpha \neq \alpha)$

as 'the possibility for α of being two or more things' ([3], p. 170) is not much comfort. For (9) then reads: the possibility for x of being less than two objects *as well as* being two or more objects.

One tack which we might think would avoid this problem in fact does not: we might try replacing the existential counterpart quantifiers (CQs) in clause (a) of MCCT with universal CQs. But the problem remains because the possibility operator in (9) continues to introduce existential CQs (for each occurrence of 'x') that have wider scope than the CQs introduced by clause (a).[4]

I can see no obvious remedy. Removing the likes of (9) would surely result in a substantially different system, one in which (5), for example — much favoured by Forbes — is likely to come out unsatisfiable.

I submit, then, that CT* remains a less problematic alternative to canonical counterpart theory.[5]

Trinity College,
Dublin 2, Eire

REFERENCES

[1] Graeme Forbes, 'Canonical Counterpart Theory', *Analysis* 42 (1982) 33–7.

[2] Graeme Forbes, *The Metaphysics of Modality* (Oxford: Oxford University Press, 1985).

[3] Graeme Forbes, 'Counterparts, Logic and Metaphysics: Reply to Ramachandran', *Analysis*, this issue.

[4] David Lewis, *Counterfactuals* (Oxford: Basil Blackwell, 1973).

[5] Murali Ramachandran, 'An Alternative Translation Scheme for Counterpart Theory', *Analysis* 49 (1989) 131–41.

[6] Murali Ramachandran, 'Contingent Identity in Counterpart Theory', *Analysis*, this issue.

---

[3] (9) gets translated on MCCT's (and RMCCT's) scheme as

$$\exists w \exists x_1 \ldots \exists x_4 (Cx_1xw \;\&\; \ldots \;\&\; Cx_4xw \;\&\; \exists y \exists z (Cyx_1w \;\&\; Czx_2w \;\&\; (y=z)) \\ \&\; {\sim}\exists y \exists z (Cyx_3w \;\&\; Czx_4w \;\&\; (y=z)))$$

[4] (9) now gets translated as:

$$\exists w \exists x_1 \ldots \exists x_4 (Cx_1xw \;\&\; \ldots \;\&\; Cx_4xw \;\&\; \forall y \forall z ((Cyx_1w \;\&\; Czx_2w) \supset (y=z)) \\ \&\; {\sim}\forall y \forall z ((Cyx_3w \;\&\; Czx_4w) \supset (y=z)))$$

[5] My thanks to Graeme Forbes and Peter Smith for comments on an earlier draft.

TRUTH THEORETICAL SEMANTICS AND AMBIGUITY

By BRENDAN S. GILLON

DAVIDSON has proposed that a semantic theory for a natural language should emulate truth conditional semantic theories for artificial languages, first developed by Tarski [10]. At least two philosophers, Kathryn Parsons [7] and L. Jonathan Cohen [1], think that his proposal will founder on the phenomenon of lexical ambiguity. While each shows that Davidson's approach to this phenomenon is inconsistent with the spirit, if not the letter, of his proposal, neither shows that every approach is. In fact, as I shall show below, there is a straightforward treatment of lexical ambiguity which follows, virtually directly, from Davidson's own explicitly stated assumptions.

Recall that a truth definition for a language is recursive and that its recursion depends on the recursive enumeration of the language's syntactic entities. The syntactic entities, as Davidson ([2], p. 311) clearly states, are structural descriptions, which, for the sake of simplifying the discussion here, are assumed to be phrase markers. Phrase markers, then, are the entities to which semantic values are assigned.

A phrase marker, however, is not the same as an expression. An expression is a phonic or graphic form: the former has acoustic properties and the latter spatial. (For further elaboration, see Lyons [5], chapter 3.3.) A phrase marker has neither acoustic nor spatial properties: it is an abstract syntactic entity. The two are related through a language's phonological rules, whereby each phrase marker is mapped onto a phonic expression. (See Selkirk [9] for how this mapping might be worked out within a generative linguistic framework.) In addition, a phonic expression is mapped onto a graphic one by rules of phonetic transcription (unless, of course, the orthography of the language is ideographic, and not phonetic).

The fact is that often more than one phrase marker is mapped onto the same expression. When this happens, there is ambiguity. Usually, if ambiguity is described in this way, one thinks of phrasal ambiguity. Parsons does not discuss phrasal ambiguity; and while Cohen ([1], p. 129) mentions it, he says that it 'creates no special difficulties'. However, he does not say why. But, in fact, it is instructive to see why.

Consider, to begin with, Lycan's ([6], chapter 2.2) compact version of the argument Parsons [7] adduced to show the problem posed by lexical ambiguity for truth theoretical semantics. A Tarski-like truth definition for a language L will, among other things, entail, for every predicate P of language L, a biconditional of the form:

(1) For any x, R is true of x in L iff x Q

(where R is a structural description of the predicate P of language L and Q is its translation into the meta-language of L). Now, let L be English and let P be 'is a bank'. Then the schema in (1) instantiates into the following assertions:

(2.1) For any x, 'is a bank' is true of x in English iff x is a financial institution.

(2.2) For any x, 'is a bank' is true of x in English iff x is a side of a river.

These two assertions yield the clearly false consequence that

(3) For any x, x is a side of a river iff x is a financial institution.

Now consider the same argument adapted to a case of phrasal ambiguity:

(4) Each old man and woman left.

Presumably, it would be a consequence of a truth definition for English that

(5.1) 'Each old man and woman left' is true in English iff each old man left and each woman left.
(5.2) 'Each old man and woman left' is true in English iff each old man left and each old woman left.

But this yields the false consequence that

(6) Each old man left and each woman left iff each old man left and each old woman left.

Clearly, this latter argument is fallacious. The source of the fallacy is the inadequate notation used to pick out the phrase markers needed in (5). (For an illuminating discussion of the many uses of quotation marks, see Lyons [5], chapter 1.2 and 1.3.) The inadequacy is obviated by the use of any standard linguistic notation. Since the phrase marker [S[NP[DET each][N[AP old][N[N man][CONJ and][N woman]]]][VP left]] is distinct from the phrase marker [S[NP[DET each][N[N[AP old][N man]][CONJ and][N woman]]][VP left]], the false consequence in (6) does not follow from the two premisses in (5).

Now, if one treats lexical ambiguity as a special case of phrasal ambiguity, then the problem raised by Parsons disappears. And it is easy to do this. A phrase marker consists of a finite set of partially ordered nodes one of which is distinguished as the greatest; its non-terminal nodes are assigned syntactic categories, together with other syntactic features; and its terminal nodes are assigned elements each of which comprises the address of a lexical entry and the syntactic properties associated with it, such as its (lexical) syntactic category and its subcategorization frame, if there is any. Moreover, it is recursively defined, so that a sub-phrase marker is itself a phrase marker. (See Higginbotham [4].) This implies that the labelled terminal nodes of any non-trivial phrase marker are trivial phrase markers. To distinguish these syntactic entities from the phonic forms onto which they are mapped by the phonological rules of the language, I shall enclose their graphic forms in angle brackets. This simple revision in standard linguistic notation enables one to see that Parsons' argument pertaining to lexical ambiguity suffers from the same fallacy as the argument just constructed for phrasal ambiguity, for the phrase marker of the predicate in (2.1), namely, [VP[V⟨is⟩][NP[DET⟨a⟩][NP⟨bank⟩$_1$]]], is distinct from the one in (2.2), namely, [VP[V⟨is⟩][NP[DET⟨a⟩] [NP⟨bank⟩$_2$]]], since the node with the label '⟨bank⟩$_1$' is distinct from the node with the label '⟨bank⟩$_2$'.

The nub of the solution is to treat lexical ambiguity as a special case of phrasal ambiguity. A definition to implement this can be easily framed.

(7) An expression is ambiguous iff more than one phrase marker is mapped into it.

In the case where the ambiguous expression is either a phrase or a sentence, the phrase markers mapped onto it are non-trivial, each consisting of more than one node; in the case where the ambiguous expression is a word, the phrase markers mapped onto it are trivial, each consisting of a single node, which includes, among other things, the lexical address of a lexical entry. Standard lexicographical practice, which Davidson ([3], p. 180) is prepared to countenance, guarantees that ambiguous words have distinct lexical addresses.

The problem pointed out by Cohen ([1]) is also overcome. He observes that natural languages, like English, contain a rule whereby a word whose literal denotation is some thing, can also denote a representation of that thing, and that this rule applies recursively. For example, Cohen ([1], p. 133) says, 'the word "face", say, can not only come to mean – in appropriate circumstances – a representation of a face but also a representation of a representation of a face, and so on, as becomes apparent immediately one considers what one could say about a picture of a picture gallery in which one of the pictures is also a picture of a picture gallery.' As a result, Davidson's ([3], p. 181) treatment of lexical ambiguity will require that the truth conditional clauses of the meta-linguistic assertion satisfying the schema in (1) be non-finitary – clearly an unwelcome result.

However, this phenomenon is no problem for the treatment of lexical ambiguity advocated here. As linguists have long recognized, some lexical entries are derived from others. The rules relating these derivations are called 'lexical rules'. (See Selkirk [8] for discussion.) Such rules relate, for example, nouns with their corresponding verbs, such as the noun 'surface' with the verb 'surface'. There is no reason why these rules cannot be recursive and provide the derived lexical entries, even an infinite number of them, with unique lexical addresses.

This treatment of lexical ambiguity obviates still another problem facing Davidson's [3] approach, not observed by either Parsons or Cohen. Ambiguity often permits sentences, for a fixed state of affairs, to be both truly affirmed and truly denied. Indeed, this fact constitutes a test for ambiguity. (See Zwicky and Sadock [11] for relevant discussion.) On Davidson's approach, if a sentence containing a word which is ambiguous, say two ways, then its meta-linguistic assertion stating its truth conditions contains a pair of disjuncts corresponding to each of the two construals. If, for a fixed state of affairs, the sentence can be truly affirmed, then at least one of the disjuncts holds; and if, for the same fixed state of affairs, it can be truly denied, then neither of them holds. But the semantic theory prohibits this, on pain of its inconsistency. In other words, the test for ambiguity is inconsistent with the demands of the truth definition. So, the disjunctive treatment of ambiguity is empirically inadequate.

But now, the fact that ambiguity often permits a sentence to be both truly affirmed and truly denied of a fixed state of affairs is no longer problematic. After all, truth is a property primarily of phrase markers and only derivatively of the sentences which express them. The same sentence can be both truly affirmed and truly denied of the fixed state of affairs because at least two distinct phrase markers, one of which the state of affairs makes true and the other of which it makes false, are mapped onto the same sentence.

University of Toronto,
Toronto, Ontario, Canada M5S 1A1

REFERENCES

[1] L. Jonathan Cohen, 'A Problem about Ambiguity in Truth-Theoretical Semantics', *Analysis* 45 (1985) 129–34.
[2] Donald Davidson, 'Truth and Meaning', *Synthèse* 17 (1967) 304–23.
[3] Donald Davidson, 'Semantics for Natural Language', in *Linguaggi nella Società e nella Tecnica*, edited by B. Visentini *et al.* (Milano: Edizioni de Communità, 1970), 177–88.
[4] James Higginbotham, 'On Semantics', *Linguistic Inquiry* 16.4 (1985) 547–94.
[5] John Lyons, *Semantics*, 2 vols. (Cambridge: Cambridge University Press, 1977).
[6] William G. Lycan, *Logical Form in Natural Language* (Cambridge, Massachusetts: MIT Press, 1984).
[7] Kathryn Pyne Parsons, 'Ambiguity and the Truth Definition', *Noûs* 7 (1973) 379–94.
[8] Elisabeth O. Selkirk, *The Syntax of Words* (Cambridge, Massachusetts: MIT Press, 1982).
[9] Elisabeth O. Selkirk, *Phonology and Syntax: The Relation between Sound and Structure* (Cambridge, Massachusetts: MIT Press, 1984).

[10] Alfred Tarski 'The Concept of Truth in Formalized Languages', in *Logic, Semantics, and Metamathematics*, edited by J. H. Woodger (Oxford: Oxford University Press, 1956), 152–278.

[11] Arnold M. Zwicky and Jerrold M. Sadock, 'Ambiguity Tests and How to Fail Them', *Syntax and Semantics* 4 (1975) 1–36.

FITCH AND INTUITIONISTIC KNOWABILITY

By PHILIP PERCIVAL

AN IDEA underlying verificationism and its contemporary development, anti-realism, is that truth is constrained epistemically. This idea has received various expressions, but increasingly prominent among them is the supposition that all truths can be known. What this amounts to depends on how 'can be known' is interpreted. One alternative, suggested by Dummett's anti-realism about the past, tenses 'can' so that 'can be known' means 'known in a possible *future*'. But on a weaker interpretation 'can be known' means 'is known *sometime* in a possible *world*'. I will call the idea that all truths can be known under this weak interpretation the 'principle of knowability' and anyone who espouses it a 'knowability theorist'.

The knowability principle appears formalizable as $p \rightarrow \Diamond Kp$, with '$\rightarrow$' read as 'if . . . then' and '$\Diamond$' and 'K' interpreted as 'it is metaphysically possible that' and 'it is known sometime that'. Appealing to a proof in Fitch [3], Hart [4] contends that this schema is absurd because it entails $p \rightarrow Kp$. The general idea of the proof is simple. If all truths are knowable, a true conjunction of the form $p \,\&\, \neg Kp$ must be knowable too. But no-one *could* know a proposition of this form, since knowing the first conjunct would invalidate the second. So if all truths *are* knowable there are no true propositions of the form $p \,\&\, \neg Kp$. And in that case $p \rightarrow Kp$.[1]

Although Hart thinks $p \rightarrow Kp$ absurd, he observes that it appears valid under an intuitionistic interpretation of '$\rightarrow$'. This reflection is curious because it suggests a retort he doesn't assess. Dummett has repeatedly argued, on grounds independent of Fitch's proof, that anti-realism warrants intuitionistic rather than classical logic and an assertibility-conditional rather than a truth-conditional theory of meaning. So Hart should have anticipated the knowability theorist responding to his charge of absurdity by arguing that while $p \rightarrow \Diamond Kp$ and its intuitionistic consequences must be understood non-classically they are unexceptionable when so construed. I will call this response the 'intuitionistic defence'.[2]

---

[1] The details of the proof are treated rigorously in Williamson [8]. Aside from its reliance on *classical* logic, the only principle on which the proof relies which might be questioned is distributivity of knowledge over conjunction: $K(p \,\&\, q) \vdash Kp \,\&\, Kq$. Williamson [11] argues that there is no prospect of defending the claim that all truths are knowable by disputing this principle.

If, as this defence suggests, classical propositional logic is relinquished in favour of its intuitionistic fragment (IPL), Fitch's proof stops short at $p \rightarrow \neg\neg Kp$, the intuitionistically invalid rule of double negation elimination being required for the additional step to $p \rightarrow Kp$. But this schema is still problematic, and an intuitionistic defence of the knowability principle owes us an explanation of why, appearances to the contrary, we should accept it. In its strongest form a defence of this kind must furnish an independently plausible semantics for the connectives validating both IPL and $p \rightarrow \neg\neg Kp$.

Published defences of $p \rightarrow \Diamond Kp$ along these lines fail to meet this standard. Rasmussen and Ravnkilde ([6], footnote 77) argue that Fitch's proof is inconsequential because (even) $p \rightarrow Kp$ is valid under a proof-conditional interpretation of '$\rightarrow$', and although Williamson [8] denies this he reaches a similar conclusion concerning the proof by arguing that $p \rightarrow \neg\neg Kp$ *is* valid under it. But this dispute is really beside the point. As anti-realist sympathizers like Dummett and Wright now admit, non-mathematical statements aren't susceptible to *proof*, and a proof-conditional interpretation of '$\rightarrow$' isn't *generally* viable. So an intuitionistic defence can't appeal to it. Williamson [10] recognizes as much, but he fails to consider alternative non-classical semantic proposals.

Williamson's omission is understandable because anti-realists have tended to devote more time to decrying truth-conditional semantics than they have to developing semantic theories of their own. Indeed, Williamson's ([10], p. 432) intuitionistic defence is grounded on the presumption that 'if a constructivist treatment of the semantics of defeasible statements can be given, we certainly do not know enough about it to defeat the expectation that it will allow (an intuitionistic defence)'. But I will argue that Williamson is wrong about this. Firstly, I will give two reasons for thinking that $p \rightarrow \neg\neg Kp$ ought *not* to be validated by *any* semantics validating an extension of IPL. (The first reason shows that it isn't *logically* valid, the second that it is untrue.) Secondly, I will argue that those semantic proposals which *are* currently available to the knowability theorist don't validate an extension of IPL including $p \rightarrow \neg\neg Kp$.

---

[2] In arguing that the knowability principle should be formalized not as $p \rightarrow \Diamond Kp$ but as $Ap \rightarrow \Diamond KAp$, with 'A' being read as 'actually', Edgington [2] develops an alternative response. However, Williamson [9] and Percival [5] show that her proposal is unsatisfactory.

In IPL, $\neg\neg\neg p \leftrightarrow \neg p$. So if the knowability theorist's extension of IPL includes $p \rightarrow \neg\neg Kp$ it will also include $\neg Kp \rightarrow \neg p$. Williamson [10] defends this schema against my charge that it is absurd by focusing on the mathematical case and arguing that it holds under a constructivist interpretation of '$\rightarrow$' because the only possible ground for supposing a mathematical proposition never known would be a proof of that proposition's negation. This limited case has some plausibility, even if we are left wondering how $\neg Kp \rightarrow \neg p$ might be justified for non-mathematical propositions. But my ground for supposing that no acceptable semantics extending IPL can validate $p \rightarrow \neg\neg Kp$ casts doubt even on Williamson's idea that this schema might be *logically* true when restricted to mathematical propositions. It is this. Since knowledge implies truth, $\neg p \rightarrow \neg Kp$ is a theorem and supposing $\neg Kp \rightarrow \neg p$ logically true means that $\neg Kp$ and $\neg p$ are logically equivalent. *But this cannot be.* For even if $\neg p$ is mathematical, and, hence, non-contingent, $\neg Kp$ is contingent unless p is false. For example, while Fermat's last theorem is non-contingent, the proposition that it is never known is contingent unless the theorem itself is false.[3]

While the fact that $\neg Kp$ is contingent even if p is non-contingently true is itself damaging enough to the idea that they are logically equivalent, it also generates counterfactuals which are inconsistent with that supposition. If $\neg p$ is non-contingently false and it is known that p, we have both $\neg q \mathrel{\Box\!\!\rightarrow} \neg Kp$ and $\neg q \mathrel{\Box\!\!\rightarrow} p$ if that knowledge is dependent on the contingency q. And while it is implausible to suppose that sometime *mathematical* knowledge can depend on such a contingency, sometime knowledge of *a posteriori* necessary truths can do so. For example, imagine a laboratory researcher about to record the litter of Doris rabbit who turns around to see that Mabel rabbit shifts her offspring Peter into

---

[3] Admittedly, one cannot conclude that $\neg p$ and $\neg Kp$ can't be logically equivalent from the fact that they express propositions having a different modal status *too* quickly. Whereas logical equivalence is a property of *types*, contingency and necessity are properties of *propositions*. And the fact that the two types are logically equivalent does not imply that they express the same proposition. Nor, even, does it imply that the propositions they express have the same modal status. For example, the types S and $(\iota x)(x=a \,\&\, S)=a$ are logically equivalent in the natural non-Fregean semantics for a primitive description-operator 'ι' (see Taylor [7]). But they don't express the same proposition if this operator is used referringly so as to contribute a *particular* object to the proposition expressed. Still, it would be wrong to invoke this kind of consideration in favour of the idea that $\neg Kp$ and $\neg p$ are logically equivalent. In the case just cited the fact that two types express different propositions while being logically equivalent is explicable because what proposition $(\iota x)(x=a \,\&\, S)=a$ expresses read de re depends on whether S is true. So to speak, in models in which S is true $(\iota x)(x=a \,\&\, S)=a$ expresses a necessary truth, while in models in which S is false it expresses no proposition at all. That is why there are no models in which S is true and $(\iota x)(x=a \,\&\, S)=a$ false or vice-versa. But there is no reason to suppose that $\neg Kp$ and $\neg p$ participate in this kind of relationship. What proposition $\neg p$ expresses under an arbitrary interpretation does not depend on the truth-value of $\neg Kp$ under it.

Doris's nest. He then thinks: if I hadn't looked around it would never have been known that Mabel is Peter's parent.[4]

If the trivial consequence of $p \rightarrow \neg\neg Kp$ just noted gives reason to think that it isn't *logically* valid, a second one shows that it isn't even true. Two applications of modus ponens to $\neg Kp \rightarrow \neg p$ gives the contradiction $\neg p \,\&\, \neg\neg p$ from $\neg Kp \,\&\, \neg K\neg p$. So $\neg(\neg Kp \,\&\, \neg K\neg p)$ is a theorem, one saying that no statement remains forever undecided. But this is unacceptable since, for some values of p, $\neg Kp \,\&\, \neg K\neg p$ is as plausible as *any* contingent conjunction of negated statements. I have just tossed a coin from my pocket ten times without checking any of the results. So neither I nor anyone else will ever know whether or not heads landed five times. Consequently, $\neg(\neg Kp \,\&\, \neg K\neg p)$ isn't generally true, and nor can it be justifiably believed. So, granted IPL, Fitch's proof shows that $p \rightarrow \Diamond Kp$ isn't generally true either and cannot be justifiably believed.

Since Fitch's proof shows that $p \rightarrow \Diamond Kp$ has consequences that are plainly unacceptable we *do* know in advance that no intuitionistic defence of the principle of knowability fleshed out with a specific semantics for the logical connectives is going to work. But I also claimed that no extant semantic proposals attractive to the intuitionist *do* validate the problematic schema Fitch derives. Since this provides further ammunition against such defences it is worth reviewing why they fail to validate the schema in question.

Dummett's arguments for the generalization of intuitionistic ideas to non-mathematical discourse have led to just two semantic proposals since the rightful abandonment of the idea that the semantics of non-mathematical statements might be effected in terms of conclusive verification-conditions. Since neither of them attempts a semantics of K it is slightly awkward to consider whether they validate $p \rightarrow \neg\neg Kp$. But provided we take for granted an informal interpretation of K we can still investigate, albeit less surely, which theorems arise when they are extended to deal with it. For example, if we are trying to decide whether $p \rightarrow \neg\neg Kp$ is validated by a semantics which extends a semantics for '$\rightarrow$' and '$\neg$' given in terms of Q-conditions, we still have recourse to our intuitions about the Q-conditions for 'Kp' even if no such extension is given formally.

Bearing these limitations in mind we can turn to the semantic proposals just mentioned. Both of them trade heavily on classical truth-conditional semantics. The first maintains that the knowability theorist can embrace a truth-conditional semantics, without further ado, so long as the logic of the meta-language in which the semantics is couched is intuitionistic.[5] Clearly, this proposal doesn't disarm Fitch's proof. Simply changing the logic of the meta-language does nothing to encourage the idea that a truth-conditional semantics for '$\rightarrow$' and '$\neg$' validates $\neg(\neg Kp \,\&\, \neg K\neg p)$. Nor does changing it explain how $\neg Kp$ can be logically equivalent to $\neg p$ even though it is contingent if $\neg p$ is non-contingently false.

---

[4] Again, the logical equivalence of r and p *might* be reconciled with the counterfactuals $\neg q \mathrel{\Box\!\!\rightarrow} \neg p$ and $\neg q \mathrel{\Box\!\!\rightarrow} r$ if we could suppose that the proposition r expresses in a counterfactual situation depends on the truth-value of p in that situation. But, as before, the proposition expressed by 'Peter is Mabel's offspring' does not depend on the truth-value of 'K (Peter is Mabel's offspring)' in this way.

The approach in Wright ([12], chapter 9) is more sophisticated. Wright suggests interpreting occurrences of 'is true' in classical truth-conditional semantics as 'is superassertible', where, basically, a sentence is superassertible iff 'it is both possible, under favourable circumstances, to achieve a body of knowledge which warrants its assertion and impossible, no matter how favourable the circumstances, to enlarge on that body of knowledge in such a way that the statement ceases to be warrantedly assertible'.

Again, $\neg(\neg Kp \,\&\, \neg K\neg p)$ isn't validated on this account since some instances of it are *not* superassertible. Mirroring the classical clause, Wright's clause for '$\neg$' has '$\neg p$' superassertible iff $\neg$(p is superassertible). So $\neg(\neg Kp \,\&\, \neg K\neg p)$ is only validated if we can always deny on semantic grounds that $\neg Kp \,\&\, \neg K\neg p$ is superassertible. This means being in a position to deny that both Kp and $K\neg p$ are not superassertible. But for some values of p we have reasons to hold neither Kp nor $K\neg p$ superassertible that are no less strong than *any* claim to the effect that such and such a proposition is not superassertible. For example, if it is not superassertible that Henry VIII never married then it is not superassertible that it is sometime known whether or not the coin tossing to which I alluded earlier resulted in five heads.[6]

It is important to dwell on the significance of these results. One route to $p \rightarrow \Diamond Kp$ contends that we couldn't grasp a truth-*condition* it might prove impossible for us to know obtained when it did. This route seems committed to a *semantic* validation of $p \rightarrow \Diamond Kp$. But if the idea that, logically, truth cannot transcend knowability has some initial plausibility, the ideas that, logically, every proposition never known is false ($\neg Kp \rightarrow \neg p$), and that, logically, no proposition will remain forever undecided $\neg(\neg Kp$ &

---

[5] This proposal relies on the idea that a truth-conditional theory of meaning only presupposes a verification-transcendent conception of truth if the logic of its meta-language is classical.

[6] Against this one might try to argue that on a strong temporal reading of 'possible' in Wright's characterization of 'superassertibility', namely 'in a possible future', it is not superassertible that the number of heads resulting from my coin toss is n = five, and, hence that the negation of this proposition is superassertible. One might then contend that it is wrong after all to suppose, as I have, that this proposition is forever unknown. But it is unacceptable for a semantics to tell me that I know that the number of heads resulting from my coin toss was neither 1 nor 2 nor . . . 10. Such reasoning only shows that either the temporal reading of 'possible' should be weakened to 'in a possible world' or Wright's clause for negation should be abandoned.

$\neg K\neg p$), are quite unattractive. And as if these consequences aren't destructive enough Fitch's proof also shows that supposing $p \rightarrow \Diamond Kp$ logically true commits one to supposing logically equivalent two types whose modal status is different ($\neg Kp \leftrightarrow \neg p$). Nor does retreating from the idea that the knowability principle is semantically validated to the idea that it is *simply* (non-logically) true, or justifiably believed or whatever, help matters. Motivation for the knowability principle has always come from the philosophy of language, and unless its justification comes from semantic theory its appeal is lost. And in any case the idea that $\neg(\neg Kp \;\&\; \neg K\neg p)$ is generally true, or justifiably believable, is just unacceptable. If we bought in to the thesis that truth can't transcend knowability we had no idea it would cost us the triviality that some propositions remain forever undecided. Of course, advocates of the principle of knowability typically view it as revisionary of *some* aspect of our linguistic practice, just as Dummett argues to non-classical logic. But if there were no limits on the revision of our opinions argument would be pointless. Maybe it is plausible to reject classical logic on philosophical grounds. But, in the style of Appiah [1], I'm suggesting that we have crossed the bounds of plausibility by the time we reach the consequences of $p \rightarrow \Diamond Kp$ Fitch's proof identifies.

16 Aylsham Road,
Norwich NR3 3HG

REFERENCES

[1] A. Appiah, 'Anti-realism unrealised', *Philosophical Quarterly* 34 (1984) 85–103.
[2] D. Edgington, 'The Paradox of Knowability', *Mind* 94 (1985) 557–68.
[3] F. B. Fitch, 'A Logical Analysis of some Value Concepts', *Journal of Symbolic Logic* 28 (1963) 113–18.
[4] W. D. Hart, 'The Epistemology of Abstract Objects', *Proceedings of the Aristotelian Society* Supp. Vol. 53 (1979) 153–65.
[5] P. Percival, 'Knowability, actuality and the metaphysics of context-dependence', forthcoming in *Australasian Journal of Philosophy.*
[6] S. A. Rasmussen and J. Ravnkilde, 'Realism and Logic', *Synthese* 52 (1982) 379–437.
[7] B. Taylor, *Modes of Occurrence* (Oxford: Blackwell, 1985).
[8] T. Williamson, 'Intuitionism Disproved?', *Analysis* 42 (1982) 203–7.
[9] T. Williamson, 'On the Paradox of Knowability', *Mind* 96 (1987) 256–61.
[10] T. Williamson, 'Knowability and Constructivism', *Philosophical Quarterly* 38 (1988) 422–32.
[11] T. Williamson, 'Does knowledge distribute over conjunction', forthcoming in J. Copeland (ed.), The Arthur Prior Memorial Volume.
[12] C. Wright, *Realism, Meaning and Truth* (Oxford: Blackwell, 1987).

TWO USES OF 'KNOW'

By D. S. CLARKE, JR.

CRITICAL appraisal of Alvin Goldman's reliabilism has been directly mainly towards its aspect as a theory of justified belief. Ignored in the process has been a basic error committed by Goldman in his attempt to state an additional condition for knowledge in order to account for a certain type of Gettier-type counterexample to the classical definition of knowledge as justified true belief. The error, which occurs in both the early formulation in 'Discrimination and Perceptual Knowledge' (*Journal of Philosophy* 73, 1976, 771–91) and the subsequent version in *Epistemology and Cognition* (Cambridge, MA: Harvard University Press, 1986), Section 3.2, arises from a failure to distinguish two very different senses in which a person can be said to have knowledge. Though only Goldman's theory is discussed, it should be clear that my criticisms have applications to other reliabilist theories, e.g. Nozick's "tracking" theory, and, more generally, to attempts in recent epistemology to isolate an informational sense of 'know' for which necessary and sufficient conditions can be stated.

First, a quick review of Goldman's account. His examples are of persons who correctly identify objects on the basis of how they look to them, but lack the ability to discriminate between these objects and relevant alternatives of which they are unaware. For example, a subject S may be driving along a highway, spot what looks like a barn, and assert 'That is a barn'. S's belief that the building is a barn is justified, contends Goldman, since it is based on direct perceptual evidence. Let us suppose it is also true: the object is in fact a barn. But suppose that S is unaware of the fact that in the area in which he was driving papier mâché copies of barns were commonly placed along the highway and that if such a copy had happened to be in front of him he would also have asserted 'That is a barn'. Then we could not claim, 'S knows that the building in front of him is a barn'. He has justified true belief, but since he lacks the capacity to distinguish barns from look-alike copies and it was simply his good luck that he saw the genuine article and not the copy, he lacks knowledge. More generally, any sentence of the form 'S knows that *p*' will express a false proposition, according to Goldman, if S lacks the capacity to distinguish the state of affairs described by *p* from relevantly alternative states of affairs, even though S's belief in *p* may be justified and *p* true. Lacking the appropriate capacity, S is an unreliable informant when he or she asserts that *p* is the case. Since the truth of *p* on a given occasion will be a matter of accident or good luck, the typical common element in Gettier-type examples, S lacks knowledge.

Goldman's condition of discriminative capacity does seem applicable to our use of 'know' when applied to a potential source of information and we are determining whether we can rely on this source for accuracy. Thus we ask, 'Does S know what time it is?', 'Does S know the length of the boat (the screw, the strand of DNA molecule)?', 'Does S know the colour of the flower (the wall in the house)?', Does S know the difference between a warbler and a vireo?', and (with Goldman) 'Does S know the difference between a barn and a barn copy?'. Here the questioner is attempting to find out whether S is a reliable source of information with regard to the type of attribute or kind of thing in question. If asked questions such as 'What time is it?', 'What is the boat's length?', or 'What is that structure?', would S provide a true answer? An affirmative answer such as 'S knows the boat's length' or 'S knows the difference between a barn and a barn copy' does imply the fulfillment of Goldman's reliability condition.

But it is important to note that this source-directed use of 'know' is not a propositional use, for it is invariably indeterminate, indicating only the general type of information that the subject S could reliably provide if asked. We use the variable terms 'length' and 'difference' in stating 'S knows the length of the boat' or 'S knows the difference between a barn and a copy', not the determinate propositional forms 'S knows that the boat is 10 metres' or 'S knows that this object is a barn'. Indeed, to use these latter sentences is to render unnecessary any source-directed use of 'know', for this use would show we already possess the information about which our potential source might be asked.

The propositional forms are instead used as answers to questions designed to find out whether a person is a potential *recipient* (not source) of information, whether information that could be provided would be redundant. Thus, we ask 'Does S know that the road is blocked?' as a means of determining whether the sentence 'The road to town is blocked' will convey useful information to S or whether it will simply repeat what he has already heard or read. In effect, the question is whether it is necessary to tell S about the road. A negative answer to the question 'Does S know that *p*?' is normally followed by sharing with S the information that *p* is the case; a positive (e.g. 'S knows the road is blocked', 'S knows that the boat is 10 metres', 'S knows that the object before him is a barn') leads normally to a decision to withhold information to avoid redundancy.

It is such answers that Goldman, in common with the classical tradition, claims to be analysing. Yet it is obvious that a true answer does not require the satisfaction of his reliability condition. The redundancy avoidance use of 'know' requires only that the agent have true belief. How he arrived at the belief, whether it is justified, whether he himself is a reliable informant — all of this is irrelevant to the question whether he needs to be informed about the state of affairs being described. Tom correctly informs Bill, 'That object is a barn', without being aware of the presence of barn-copies. If Bill were then to say to him 'That is a barn', Tom's natural (and perfectly justified) response would be 'I know that'. Tom doesn't know that there are barn-copies in the area; he does require information about this state of affairs. He may not know, when provided this information, the difference between a barn and a copy; in this case he would have to learn to make the relevant discriminations. But he does certainly know in the only relevant sense of 'know' that the object before him is a barn.

Goldman's mistake is to extend a condition (indeed, the sole condition) for the source-directed non-propositional use of 'know' to its propositional redundancy avoidance use. Why does he commit it? The reason may be that there is a connection between the uses that tends to lead us to confuse one for the other. If we know that someone is a source of reliable information about a general subject matter, we think it would be redundant to provide him propositional information about what he is presumably already informed. Thus, if S knows the length of the boat or the difference between a barn and a relevant alternative, we infer he knows that the boat is 10 metres long, that this object is a barn and not a copy. In this sense non-propositional source knowledge is *sufficient* for propositional redundancy avoidance knowledge. This may lead us then to conclude that the converse holds, that source knowledge is also *necessary* for redundancy avoidance knowledge. And this is precisely the error.

Southern Illinois University at Carbondale,
Carbondale, IL 62901, U.S.A.

INTENDING AND MOTIVATION

By JOSEPH MENDOLA

MANY writers on intention claim that someone x intends to A only if x is preponderantly motivated to A, more motivated to A than to do something incompatible[1] ([1], [2] pp. 304–305, [3], [6]). In his book *Irrationality* and in 'Intending and the Balance of Motivation', Alfred Mele contends that this popular thesis is false ([4], [5] pp. 31–49). His case rests on what friends of the thesis will hold to be two sorts of examples, and my claim is that the examples do not establish his contention.

---

[1] Some of these theorists qualify the claim in important ways which we will ignore in this discussion. For instance, it may be that someone may intend to do something incompatible with something else they are more motivated to do, when they fail to realize that the two things are incompatible.

The first kind of example Mele cites consists of cases of *akrasia* such as the following: Alex and Bob swear to become blood brothers, and Alex, after the appropriate practical reasoning, forms the intention to cut his hand here and now with a knife in the requisite way. He brings the knife down toward his palm to do so, only to find himself unable to proceed, defeated by his motivational condition ([4], p. 371).

On Mele's reading of the case, this is what happens: Before bringing down the knife, Alex forms the intention to cut his hand here and now, but his failure to go through with the act indicates he lacks preponderant motivation to do so. He does *begin* to perform the act, and this indicates he has preponderant motivation to *begin* it, but he has the intention to cut his hand while lacking preponderant motivation to do so.

There is, as Mele admits, an alternative reading of such cases, consistent with the popular thesis he denies: When Alex forms the intention and begins to bring down the knife, he is preponderantly motivated to cut his hand, but as he brings it down and gets close to the flesh, his motivation changes. We are all familiar with forming some resolution supported by what seems preponderant desire, and then finding our balance of motivation shift in the immediate presence of the relevant opportunity, and there is a developed psychological literature centring on such approach-avoidance behaviour. The fans of the popular thesis which Mele denies might insist that his case can be treated as analogous to these familiar cases, and hence that such a case cannot serve to establish that Mele's view is to be preferred to the popular thesis.

Mele cites what his opponents will hold to be another class of cases in support of his view. Consider the case of someone forming the intention not to smoke, though their preponderant motivation leans the other way ([5], p. 37). It seems that we can form intentions on the basis of practical reasoning that runs contrary to our antecedent preponderant motivation, and in fact such reasoning can influence our motivational condition. But if preponderant motivation were implied by the presence of an intention as the popular thesis holds, it is hard to see how this could be.

But Mele's opponents have a coherent story to tell about such cases as well. Let them postulate a psychological state distinct from intention, called perhaps 'the conclusion of practical reasoning', and let them admit that conclusions of practical reasoning are not necessarily in tune with preponderant motivation in the moment when they are drawn, though if reasoning is an action which is subject to motivation perhaps one has preponderant motivation to adopt such as the conclusion of practical reasoning. And then let them claim that this conclusion only becomes an intention if and when the conclusion is accompanied by supporting preponderant motivation, which it has some tendency to create, other things being equal.

Mele might insist that such a story is plausible only if the intention not to smoke, for instance, is at least weakly successfully, i.e. only when the smoker develops sufficient motivation to at least *attempt* not to smoke. He might insist that we can form an intention and lack even sufficient motivation to attempt to act on it. But his opponent may plausibly reply that it is far from obvious that under such conditions there is an *intention* not to smoke.

I should admit that fans of the popular thesis can handle the cases Mele presents only by helping themselves to particular psychological theses, for instance claims about how motivation changes over time, but this is not especially a difficulty for them. *Any* view about the connection between intention and motivation will be subject to counterexample if we allow opponents of the view to deny or assert at will any subsidiary thesis regarding how motivation changes over time. It isn't Mele's cases alone, but those cases in conjunction with subsidiary psychological theses, which could establish his contention. But those subsidiary theses are not established by his cases, nor are they obviously true on other grounds.

Mele might attempt to retreat to a weaker conclusion than I have attributed to him. He might merely claim to have shown that meaning conventions and logic do not establish that intending *entails* preponderant motivation; alternatively, he might merely claim that it is in some relevant sense *possible* that one intend to *A* and not be preponderantly motivated to *A* even if in fact only those who are preponderantly motivated to *A* intend to *A*. In fact, he explicitly claims that he need not rely on "real world" examples to support his contention ([5], pp. 37–8). But to the degree that meaning conventions regarding 'intention' are not strong enough to generate the plausible thesis Mele denies and to eliminate the possibilities he considers, it is far from clear that they generate a determinate enough inter-theoretic reference of 'intention' to allow there to be a real issue between Mele and those he attacks. This is, I think, the crucial point.

Cases fail to adjudicate the dispute between Mele and his opponents. Each side can handle all the phenomena equally well. It seems hard to see how any knowledge of the neurological structures which on Mele's view instantiate all this psychology could help decide the issue. Nor do common meaning conventions governing the use of the folk-psychological terms that Mele and his opponents employ seem to favour one side over the other. It may even be hoping too much to think that there is a determinate enough inter-theoretic reference of a term like 'intention' for that to help much, for there to be determinate states which are called 'intentions' by all parties to the dispute and which either in fact involve preponderant motivation or do not. Nor does introspection seem likely to turn the tide one way or the other. How are we to react to this seeming standoff?

Perhaps we should think that there are cases and evidence yet unconsidered or that this is an issue which our evidence might not adjudicate but which still could have a correct answer. Perhaps that answer turns on something like the most fruitful form of future intentionalist psychology. Perhaps when we complete an intentionalist psychology along Mele's lines and another along those of his opponents we will see that those psychologies attribute different causal structures to the mind and that one of them gets it wrong, or that one of them cannot be assigned intentional content in an intuitive and appropriate way. But there is another possibility which should trouble practitioners of Mele's sort of intentionalist psychology more than it does: One doesn't need to be an eliminative materialist or behaviourist, or a fan of significant indeterminacies of psychological content, to worry that Mele and his opponents have simply different notations which capture the same facts. This is only a possibility, but we should review our evidence that it is not the case.[2]

University of Nebraska,
Lincoln, NE 68588-0321, U.S.A.

REFERENCES

[1] Robert Audi, 'Intending', *Journal of Philosophy* 70 (1973) 387–403.
[2] H.-N. Castañeda, *Thinking and Doing* (Dordrecht: Reidel, 1975).
[3] W. Davis, 'A Causal Theory of Intending', *American Philosophical Quarterly* 21 (1984) 43–54.
[4] Alfred Mele, 'Intending and the Balance of Motivation', *Pacific Philosophical Quarterly* 65 (1984) 370–76.
[5] Alfred Mele, *Irrationality* (Oxford: Oxford University Press, 1987).
[6] R. Toumela, *Human Action and its Explanation* (Dordrecht: Reidel, 1977).

---

[2] I would like to thank Al Mele, Robert Audi, Michael Bratman, and the Editor for helpful remarks on an earlier version of this comment.

INTENDING AND MOTIVATION: A REJOINDER

By ALFRED R. MELE

ELSEWHERE, I have attacked the thesis that an agent intends to *A* only if she is preponderantly motivated to *A* ([1], [3], [4]). More precisely, I have challenged the contention that an agent intends at *t* to *A* at *t** (where *t** may or may not be identical with *t*) only if, at *t*, she is more motivated to *A* at *t** than she is to perform any competing action. Although attempted counterexamples figure saliently in my criticism of the view, Joseph Mendola, in [5], misleadingly suggests that my case against the thesis 'rests on' the examples. Some of my examples — including the blood brother case discussed by Mendola — are designed to illustrate the following point: Partly because evaluation or assessment of wanted items has a firmer grip upon what we intend than upon the motivational strength of our wants, our intentions can be at odds with our preponderant motivation.

This point or claim rests partially on two more fundamental theses:

(1) The motivational force of our wants (broadly construed) is not always in line with our assessment or evaluation of the "objects" of our wants — that is, of what is wanted.
(2) Many intentions are formed, in significant part, on the basis of our *assessment* of the objects of our wants.

Now, as I explain in [3] (Ch. 6), there is considerable empirical support for (1). To make a long story short, studies of delay of gratification provide excellent evidence both that representations of wanted items have two important functional dimensions — an informational and a motivational one — and that the dimensions are not always aligned.

Walter Mischel and his colleagues, in a series of ingenious studies spanning some fifteen years, kept one feature of the experimental design constant: Children who ranked one snack higher than another were told that they could have the lower ranked snack upon request but must wait for the experimenter to return before they could have the preferred one. The length of time that they waited for the preferred treat varied markedly, depending upon whether and how the children attended to the snacks during the waiting period. Children who attended to the very features of the snacks on the basis of which they ranked them requested the lower ranked treat much earlier than did other children under a variety of different attentional conditions. To put the experimenters' conclusion simply, attention to the consummatory features of the snacks increased the children's motivation to request the lower ranked snack, even though these very features were the basis for the children's ranking of the snacks. More simply yet, even when one's evaluative ranking of a collection of items and one's desires for those items are sensitive to the same features of the items, evaluation and motivational strength are not always responsive to those features in just the same ways, nor with matching results. Real human beings are more complicated than that. (See [3], pp. 84–90, for a documented discussion of the empirical work.)

Thesis (2) is a central element in a standard conception of practical reasoning. Practical reasoning, commonly conceived, often proceeds to a considerable extent on the basis of our *assessment* of considerations; and, accordingly, many intentions formed on the basis of practical reasoning are formed, in significant part, on the basis of these assessments.

If the motivational force of considerations can be out of line with agents' pertinent assessments, and intentions issue in significant part from assessments, why should anyone doubt that, in some cases, an agent may intend to *A* and yet not be most motivated to *A*? One might worry that if intention does not essentially incorporate preponderant motivation it cannot do its motivational work. But, as I have argued elsewhere ([1]; cf. [3], p. 104), this worry rests on a mistake. Even if an intention to *A* will issue in an intentional *A*-ing only if the agent is preponderantly motivated to *A* at the time of action, this does not entail that having preponderant motivation to *A* is required for possessing an intention to *A*. Preponderant motivation may be required for the *execution* of intentions without also being required for their *presence*.

Another worry derives from a particular interpretation of ordinary usage. It might be claimed that meaning conventions dictate that intention essentially incorporates preponderant motivation. Cases of the sort that I adduce are designed, in part, to challenge the pertinent linguistic intuitions. And there is even a story to be told about how intentions and preponderant motivational force can diverge ([3], Chs. 3 and 6) — a story involving theses (1) and (2) above. If intentions are sensitive to assessments in the way that standard conceptions of practical reasoning presuppose, and assessments can be out of line with the motivational force of assessed items, we can understand how someone who, as it happens, is not most motivated to *A* may nevertheless intend to *A*.

Of course, a philosopher can stipulate that intention, as he or she chooses to use the term, essentially incorporates preponderant motivation. What I count as an intention may then be deemed, in some instances, a distinct state (sometimes, perhaps, a 'conclusion of practical reasoning') that must be supplemented by preponderant motivation if it is to be a *bona fide* intention. It is worth asking, in this connection, whether we should be willing to make the corresponding stipulation about *decision*: An agent has just made a decision to *A* only if she is preponderantly motivated to *A*. Consider an agent, *S*, who, on the basis of practical reasoning, not only decides that it would be best, all things considered, to *A*, but also decides *to A*. Must such an agent be preponderantly motivated to *A*? Perhaps certain fears harboured by *S* are more potent than she suspects, indeed so potent that, unbeknownst to her, she is more motivated not to *A* than to *A*. If so, must we withdraw the supposition that *S* has decided to *A*? Or is it rather the case that *S* has decided to *A* even though she lacks preponderant motivation to *A*?

If, as I have argued elsewhere, the control that we have over what we decide to do is not matched by our control over the strength of our desires, we should countenance the possibility that our decisions are occasionally out of line with our preponderant motivation. And if, as is commonly and plausibly supposed, deciding to *A* is a mode of intention formation, we get the same result for intention. One can stipulate, of course, that a state counts as a decision to *A* only if the agent is preponderantly motivated to *A*. But, for reasons sketched above, this is not required to account for decision's role in the production of action; nor do I see how it can be dictated by common meaning conventions about 'decision'.

Some of the foregoing might convey the impression that, at bottom, the dispute is merely verbal. However, at least for some, the issue runs much deeper. On one influential analysis of intention, the state is reducible to a combination of belief and preponderant desire, and it is this desire component of intention that explains its capacity to initiate intentional action. I have argued in [1] that intention's motivational role rests upon its having an executive feature implicit in the idea that intentions *settle* for agents the practical question what they will do or attempt. No combination of preponderant motivation and belief is both necessary and sufficient for this feature of intention. For example, an agent who is preponderantly motivated to *A* and believes that he probably will *A*, and even that his desire to *A* will probably result in his *A*-ing, need not be settled upon *A*-ing. Consider someone deliberating about whether to make yet another attempt to quit smoking. He realizes that he is preponderantly motivated to continue smoking and believes that this motivation will probably win out, partly on the grounds that his numerous past attempts to kick the habit have been unsuccessful. Still, he regards the issue as open. He believes — truly and justifiably, we may suppose — that there is some chance that his next attempt will be successful. And he is trying to decide whether to make the effort. Such an agent is not settled upon continuing to smoke.

Now, even someone who, like me, takes intention to be irreducible to belief/desire complexes might want to insist that intending essentially incorporates preponderant motivation. One can grant that intention has an executive dimension not wholly encompassed in belief and desire and even accept (1) and (2) above, while also maintaining that intending to *A* entails being preponderantly motivated to *A*. Perhaps it will be claimed that although (a) evaluative practical reasoning does play a significant role in the formation or acquisition of many intentions and (b) our assessment of pertinent items can be out of line with motivational strength, it is only when an evaluative judgment in favour of *A*-ing is aligned with, or helps to issue in, preponderant motivation to *A* that it gives rise to something properly termed an *intention* to *A*.

Again, however, if the preponderant motivation constraint on intention is not to be merely stipulative, how is it to be defended? The constraint is not mandated, I have argued, by functional considerations, nor by meaning conventions. Moreover, as I have briefly indicated, an alternative conception of intention — one that does not make preponderant motivation an essential feature of the state — can account for intention's role in the production of intentional action. (See [2] for other functional roles of intention.) For these reasons and others, I have urged that the possibility of cases in which someone who intends to *A* is not preponderantly motivated to *A* be left open.

Davidson College,
Davidson, NC 28036, U.S.A.

REFERENCES

[1] A. Mele, 'Exciting Intentions', *Philosophical Studies* 59 (1990) 133–56.
[2] A. Mele, 'Intention, Belief, and Intentional Action', *American Philosophical Quarterly* 26 (1989) 19–30.
[3] A. Mele, *Irrationality* (Oxford: Oxford University Press, 1987).
[4] A. Mele, 'Intending and the Balance of Motivation', *Pacific Philosophical Quarterly* 65 (1984) 370–6.
[5] J. Mendola, 'Intending and Motivation', *Analysis* 50 (1990) 190–93.

WHEN DO WE DECEIVE OTHERS?

By ANNETTE BARNES

IF philosophers want to use the deception of others as the model for self-deception (and whether they should do this is a subject of continuing debate) then they must come to terms with the difficulty of explaining how people can get themselves to believe propositions which they know or truly believe to be false. This difficulty cannot be avoided, for all cases of interpersonal deception require the deceiver to get the deceived to believe something the deceiver knows or truly believes to be false. Examples in the current literature which might be thought to support a contrary view — the view that interpersonal deception and analogously self-deception require only that the deceiver persuade the deceived by 'epistemically unfair means or practices' — do not on closer inspection support this view. I will argue that other-deception *always* requires something more than some form of epistemic irresponsibility; the deceiver must also know or truly believe the proposition at issue. Hence, self-deception, if modelled on the deception of others, remains as paradoxical as many have thought.

I

It is widely assumed that if I deceive you, I must get you to believe something false. But if I get you to believe something false when I *mistakenly believe* it is true or when I *inadvertently* get you to believe it, I need not be deceiving you. Rather, to deceive you I must intentionally get you to believe something false while keeping my intention of doing this hidden from you. It seems, moreover, that the only way I can be assured of doing this is if I am in a special epistemic state (s), that is, *if I know or truly believe that a proposition I get you to believe is false.*

There are examples, however, which suggest that we find it natural to speak of deception even when epistemic state s is lacking. Frederick Schmitt, for example, argues that the deceiver *needs only to be justified in believing that something is false* ([2], p. 185). And Bas van Fraassen speaks of the deceiver as *not needing to know whether something is false.*

> Suppose that a certain bridge is dangerous, and I do not know this, that I really have no evidence that makes it less likely than not, for me, that the bridge is safe. But suppose in addition that I think it would be very good for me if you believed the bridge to be safe and I successfully persuade you to believe this falsehood. Then I have certainly deceived you, although I did not know that the information was false. ([3], p. 124)

It also seems possible that a person who *mistakenly believes that something is false* could nevertheless deceive another. Suppose I believe that your book will be published by Halcyon Press, having refereed it and recommended publication. However, I want you to believe that your book will not be published. I know, moreover, that you distrust me and will believe the reverse of what I say. By telling you that your book will be published I get you to believe that it will not be published. Much to my surprise, the press does not publish your book (another referee strongly recommended against it). In this case something I get you to believe — that your book will not be published by Halcyon Press — is true. It seems, therefore, that I cannot be in the required epistemic state s of knowing or truly believing that what I get you to believe here is false.

It has also been claimed by Brian McLaughlin that a deceiver *can even believe that what he gets the deceived to believe is true.*

> Tom believes that Dick is innocent of a certain wrong-doing. Harry believes Dick is guilty. Tom is unable to produce evidence that will change Harry's mind. So, ... Tom tries to convince Harry by lying about some particularly pertinent matter, Dick's whereabouts on a certain night ... Suppose that Tom succeeds: Harry believes the lie and infers that Dick is innocent. Then, Tom will have succeeded in intentionally deceiving Harry into believing that Dick is innocent. Yet Tom himself believed that Dick was innocent when he initiated the deceitful act. ([1], p. 35)

If, in this case, Dick was innocent I could not know or truly believe that 'Dick is innocent' was false.

It also seems possible that in a variation of the Tom, Dick and Harry case, Tom could not only believe but could know that Dick was innocent and yet deceive Harry about Dick's innocence. It could happen that Tom's good evidence does not persuade Harry of Dick's innocence but Tom's lying does.

These cases suggest that requiring the deceiver to be in epistemic state *s* is misguided. It appears that deception can take place given various epistemic states of the deceiver — he can know, truly believe, or be justified in believing that what he gets the deceived to believe is false; he cannot know whether it is false; he can mistakenly believe that it is false; he can believe or know that it is true.

II

While the described examples may at first sight seem to show that the proposed essential epistemic state *s* is absent, I will now argue they do not show this. If we attend to the following distinctions and re-examine the examples in light of them we find that the requirement that *s* be present is in fact met.

The first distinction concerns what I deceive you about. When I deceive you, I deceive you about something and there are a number of things I can deceive you about.

(1) I can deceive you about some proposition *p*.
(2) I can deceive you about what I believe about *p*.
(3) I can deceive you about what I am trying to do in the situation. For example, I might deceive you about what I am trying to get you to believe about *p*, what I am trying to accomplish in getting you to believe that *p*, what method I am using to get you to believe that *p*, and so on.

The second distinction concerns the difference between deceiving someone *into* believing something and someone being deceived *in* believing that something. Whenever I deceive you about a proposition, I deceive you *into* believing that proposition. This, as McLaughlin points out, is to be distinguished from your being deceived *in* believing that proposition. That is, I may, for example, as in the case cited earlier, by using your distrust of me, deceive you into believing that your book will not be published by Halcyon Press when I mistakenly believe it will be published. Here you are not deceived in believing that your book will not be published for your belief about its (non-)publication is true. In order for you to be deceived in believing that *p*, *p* must be false. The falsity of *p* is not required for me to deceive you into believing that *p*. However, whenever I deceive you into believing that *p*, there must be something false that I cause you to believe, you must be deceived *in* believing something. In the case under discussion, for example, you are deceived in believing that I believe, despite what I say, that your book will not be published, since I do not believe this.

I suggest that in order for me to deceive you about *p*, to deceive you into believing that *p*, the following conditions must be satisfied.

(A) I intentionally get you to believe (a) the proposition that *p* (b) a proposition about what I believe about *p* and (c) a proposition about what I am trying to do in the situation.
(B) Concerning at least one of these propositions, I know or truly believe it to be false, and intentionally get you to believe, mistakenly, that it is true.
(C) I intentionally keep hidden from you my intention of misrepresenting the truth value of the proposition or propositions in question.[1]

If we reexamine our examples in light of the two distinctions given, we shall see that the epistemic condition specified in (B) is always satisfied. While it is true, as the examples cited show, that a deceiver need not know or truly believe that *p* is false in order to deceive someone about *p*, the deceiver in these cases knows or truly believes something is false which he gets the deceiver to believe is true. The deceiver here either knows or truly believes of some proposition (*q*) about what he believes about *p*, that it is false, or of some proposition (*r*) about what he is trying to do in the situation, that it is false, which he gets the deceived to believe is true. This confirms the idea that the deceiver must be in some *s*-type epistemic state. (B) is a condition to which none of the examples mentioned serve as a counterexample.

---

[1] Cf. Brian McLaughlin's account '(B) for any *x* and for any *y*, if *x* intentionally deceives *y* into believing or continuing to believe something, *p*, then there is something, *q*, (*p* or something else) that (i) is false, that (ii) *x* disbelieves when initiating the deceitful act, and that (iii) *x* intentionally gets or allows *y* to come or continue to believe' ([1], p. 36). This is consistent with, although not equivalent to, mine.

In the book publishing example, I know or truly believe that the proposition I get you to believe about what I believe about your book's publication is false. In the Tom, Dick and Harry example, Tom knows or truly believes that he is not telling Harry the truth about Dick's whereabouts on a certain night but gets Harry to believe he is. Tom also knows that the method he is using to persuade Tom is not a reliable one but he gets Harry to believe it is.

While Brian McLaughlin does not regard his Tom, Dick and Harry case as showing the possibility of interpersonal deception, and hence self-deception, when our condition (B) is not satisfied, Bas van Fraassen does seem to regard his bridge case as showing this. As we saw, he believes that in this case, 'I have certainly deceived you, although I did not know the information was false'. I got you to believe that the bridge was safe even though I did not know or truly believe that the bridge was not safe. 'Similarly', he continues, 'I expect that there can be cases of self-deception in which the person cannot correctly be said to have known or even recognized the truth as the most likely. The main requirement for it to be (self-)deception as opposed to unintentional (self-)misleading, is that he be culpable of persuading himself by epistemically unfair means or practices.' ([3], p. 124)

Van Fraassen, is not, however, justified in concluding from this case that there is only one main requirement for (self-)deception. The person in his case knows or truly believes, for example, that he is trying to get you to believe that the bridge is safe when he does not know whether it is safe. He gets you to believe, however, that he is trying to get you to believe that the bridge is safe when he does know that it is safe.

Moreover, if we imagine that the epistemic condition specified in (B) is not satisfied, the plausibility of talking of deception vanishes.

Suppose, for example, Tom does not truly believe that what he is trying to get Harry to believe about what he, Tom, is up to is other than the truth about what he, Tom, is up to. He doesn't believe that what he is trying to do is persuade Harry by epistemically unfair means. And if Tom does not believe either that the proposition 'Dick is innocent' is false or that the proposition 'I believe that Dick is innocent', is false, there appears to be no basis for ascribing deception here. Similarly for the bridge case, *mutatis mutandis*.

III

If we use the deception of others as the model for self-deception, then as van Fraassen correctly points out, there can be some cases of self-deception in which 'the person cannot correctly be said to have known or even recognized the truth of not-p as the most likely'. However, as we have seen, in these cases the self-deceiver can correctly be said to have known the truth of some proposition not-q or not-r. The difficulty that is avoided by showing that it need not be the case that the self-deceiver knows that not-p and yet gets himself to believe that p reappears, for the self-deceiver, while knowing or truly believing not-q or not-r, must get himself to believe that q or r. Not only, as van Fraassen suggests, must the self deceiver be culpable of persuading himself by epistemically unfair means or practices, but *if* self deception is in this regard like the deception of others, then he must also get himself to believe that some proposition which he knows or truly believes is false is true.[2]

University of Maryland Baltimore County,
Baltimore, MD 21228, U.S.A.

REFERENCES

[1] Brian McLaughlin, 'Exploring the Possibility of Self-Deception in Belief', in *Perceptives on Self-Deception*, edited by Brian McLaughlin and Amélie Rorty (Berkeley: University of California Press, 1988), 29–62.

[2] Frederick F. Schmitt, 'Epistemic Dimensions of Self-Deception', in *Perspectives on Self-Deception*, 183–204.

[3] Bas C. van Fraassen, 'The Peculiar Effects of Love and Desire', in *Perspectives on Self-Deception*, 123–56.

---

[2] I would like to thank an anonymous referee and Gerald Barnes for their helpful comments on an earlier draft.

RATIONALITY IN CONVERSATION AND NEUTRALITY IN POLITICS

By GOVERT DEN HARTOGH

ONE of the aims of Charles Larmore's interesting book *Patterns of Moral Complexity* is to argue that no loss of moral substance needs to be involved, if political morality is unfastened from all conceptions of the good life.[1] Romantic critics are wont to accuse liberalism of having an abstract, bloodless, empty conception of morality. If the liberal ideas of autonomy and neutrality were meant to summarize a personal morality, then the critics would be right. But they neglect a small complicating factor: that a society like ours is characterized by a plurality of different conceptions of the good. Some of these only differ, but do not conflict. Perhaps they are not jointly realizable in a single life, but nevertheless they may be equally valuable, and accepted to be equally valuable. But often the value of incompatible conceptions, or at least their ranking, is controversial. The only way to orient political life to a comprehensive conception of the good would in that case be to compel dissenters into the outward forms of allegiance to it. The liberal political morality is best seen as an alternative way to deal with plurality and conflict: it attempts to generalize, as Rawls said,[2] the idea of religious tolerance into a social form.

---

[1] Charles E. Larmore, *Patterns of Moral Complexity* (Cambridge: Cambridge University Press, 1987). My discussion concerns chapter 3.

Political morality therefore is a morality of the second order. It is not based on a general individualistic theory of human nature, denying the importance of shared commitments, but on the experience of plurality and conflict. These ideas I heartily accept. What I want to take issue with is the thesis that such a morality of the second order has to proclaim neutrality as to first-order values, and Larmore's argumentation for this thesis.

The argument starts from the fact of disagreement about values, and it assumes, on the side of the disagreeing parties, a commitment to discuss rationally what ought to be collectively binding political principles. Let us look at a two-party discussion. Each party will propose principles adapted to its own conception of the good, and the other party will predictably reject the proposal. Now, if two parties disagree but nevertheless want to continue their conversation, they should, Larmore argues, adopt a neutral stance by setting aside, for the time being, the opinions in dispute and continuing to converse on the basis of the rest of their beliefs. In this way they could hope either eventually to resolve the dispute, or at least to bypass it. But resolution of the conflict is a forlorn hope, so the outcome of their joint commitment to the principles of rational conversation will for both parties be that they, having retreated to neutral ground, have to remain there.

Why should the parties, as this argument presupposes, want to continue their conversation? Sympathy or a sense of community could be the reason, but the minimal motive the liberal has to assume, is a shared desire for civil peace. This leads the parties to respect each other's beliefs, i.e. to see that the other's belief is justified from his perspective.

It could be objected that by restricting themselves to the overlap in their beliefs the parties may not have sufficient material to justify any principles for distributing liberties and wealth at all. In that case they should trade off the desideratum of remaining neutral against the need for some decision, but they should do so 'in the spirit of neutrality', by making the least deviation from neutrality possible, permitting into their decision making only marginal ideals, or ideals shared by the largest number.

---

[2] John Rawls, *A Theory of Justice* (Oxford: Oxford University Press, 1971), 206. I discuss Rawls' arguments for neutrality in G. A. den Hartogh, 'The Ideal behind the Exclusion of Ideals', *Rechtsfilosofie en Rechtstheorie* 18 (1989), 217–36.

This (I hope) is the skeleton of Larmore's argument. (It owes, as he acknowledges, a great deal to Habermas.) How to assess it?

When two persons are arguing about a point on which they have incompatible beliefs, then it is often reasonable for them to attempt to establish what relevant beliefs they do agree about. They both believe their own position to be the true one, and they both believe the relevant beliefs they share to be true. So, if it should be possible to argue from the common ground to one of the controversial positions, each has to assume that his own position will be the favoured one. Therefore, if they want to convince the other party, this is the obvious move to try. (And if one of them also wants to verify the truth of his own position, it is not an unreasonable move either.) It is motivated by the hope of finding assent after all.

But suppose they do not want to convince the other, or have no hope of succeeding in this? Then they will have reason to stop the conversation, or, if they must continue, to change the subject. This is a completely different move, with a different rationale: it is not a *reculer pour mieux sauter*. The maxim of polite conversation is not so much to retreat to common but rather to neutral ground: not to argue from shared beliefs, but to avoid controversial subjects.

Now if people have to find common principles of cooperation and of dividing the cooperative surplus, but disagree about the final ends of cooperation beyond all hope of consensus, which, if any, of these reasons apply to them? They have given up attempting to extend their consensus, so the first reason does not apply. But they have nevertheless to find a solution to their common problem, so they cannot simply stop discussing the matter about which they disagree and go on to other subjects. So the second reason does not apply either.

The neutrality of the liberal state is not identical with the neutrality of polite conversation. In the last case each party holds on to his position, they only agree not to discuss the matter. In the first case each party is supposed to sacrifice part of his position. For if any party had his way, his conception of the good life would be the sole foundation of the political framework. Tolerance is not the common ground between adherents of different religious views. If they do not agree whether religion A or B should be instituted as the state religion, it does not follow that they agree that neither should. Liberal neutrality is not agreeing to disagree, but accepting a compromise. A compromise is not 'neutrally justified', but differentially justified from the different perspectives of the parties concerned.

The adequate context for reasoning about moralities of the second order is not rational dialogue (about the truth of beliefs), but bargaining. If you find that the result of trying to impose your religious beliefs on others is only interminable civil conflict, then you have a good reason to agree to restricting religion to a private domain in which each adult person is his own sovereign, and which is only protected by the government, not ruled. This is not 'by-passing' a conflict, but ending it, by an exchange of concessions. Neutrality is a reasonable compromise in some cases. But it is not necessarily an optimal solution to all bargaining problems. Why should it be always impossible first to identify a neutral solution, and then going on finding outcomes which are Pareto-superior to the neutral solution? You accept a point which has a high positive value for your opponent but a low negative value to you, and vice versa.

And why, for that matter, should it always be possible to identify a neutral solution? As we saw, Larmore has his own doubts about this point. But if the case arises that we have to choose between conceptions of the good life with a necessary public dimension, this is not a choice to be made by trading off neutrality against the 'need for some decision'. (That would be a neutral desideratum itself, so we would not have to compromise neutrality for it. We could for instance have a lottery decide.) There is no collective subject, 'we', making the decision within an agreed framework of considerations. There are only individual subjects, and associations of subjects, deciding on their own accommodations within the framework of their own religious, scientific and moral views. Perhaps they agree not to use certain threats and other bargaining moves which endanger civil peace. And perhaps they agree to abide by the outcome of a certain decision procedure. These agreements are compromises themselves.

University of Amsterdam,
Nieuwe Doelenstraat 15, 1012 CP, Amsterdam, the Netherlands

THE DIVESTITURE PUZZLE DISSOLVED

By ROGER A. SHINER

STEVEN M. CAHN ([1], pp. 175–6) has offered an argument purporting to show the divestiture of shares in a company thought to be immoral cannot be required by moral principle. Many people think that such divestiture is a moral duty, and so Cahn's argument presents a challenge to such people. Cahn's argument turns on (A) the undoubted fact that Seller can only successfully sell such shares if someone else, Buyer, actually buys them, and the claims (B) that, given this undoubted fact, Seller thereby 'entangles' ([1], p. 176; [2], p. 156) Buyer in the same moral predicament from which Seller seeks to be disentangled by the divestiture, and (C) that it can never be a requirement of moral principle to entangle another in a moral predicament.

That the divestiture puzzle is not philosophically trivial is shown by Cahn's response [2] to the attempts so far published to solve it (see [3], [4], [6]). Cahn rejects all these attempts on the grounds that each of them in different ways shows how divestiture may be justifiable on grounds other than those of moral principle, grounds of a roughly consequentialist sort. If in fact divestiture cannot be justified on non-consequentialist grounds, but only on consequentialist grounds, then we would seem to have a powerful reason for rejecting non-consequentialist approaches to moral theory. That would not be a trivial matter.

(A) and (C) above seem to be unexceptionable. I shall show that (B), however, is treacherous. The term 'entangle' is too vague: if one disambiguates it, one can then show that (B) is mistaken. If (B) is mistaken, then Cahn's puzzle has no foothold.

First, another ambiguity. If a 'principled' divestiture is one that can be universalized, then we are to suppose Seller's 'I ought to sell my shares in Armaments Manufacturer' universalized to 'Everyone ought to sell their shares in Armaments Manufacturer'. In that case, since a seller requires a buyer who will then own the shares, of course principled divestiture is impossible. But it is trivially impossible, and Cahn is laughing at us behind our back. If a 'principled' divestiture is one such that Seller has a coherent non-consequentialist case for maintaining that he or she as individuals should sell their shares in Armaments Manufacturer, then the request for the form of such a case is non-trivial. I assume the problem Cahn has set is non-trivial.

There are two ways of looking at Seller's 'moral predicament'. The first concentrates on the state of Seller's soul, as it were, and regards Seller's soul as contaminated by association with the immoral company in which Seller holds shares. The second concentrates on the harm done – say, the economic support given to a corrupt regime – by the investment Seller had originally made. The 'entangling' of Buyer which purportedly takes place when Seller sells Buyer the shares might refer to either of these considerations.

Let us take the first one first. The thought here is that by the act of selling Seller purifies Seller's soul at the cost of passing on the contamination to Buyer. A certain kind of paradigm of immorality is presupposed here, cases where initially innocent persons have their moral sensibilities gradually blunted to the point where they become willing partners in some nefarious enterprise, or else where they are forced to participate in such an enterprise by blackmail or other form of coercion. Such cases require some self-conscious intention to entangle on the part of the entangler and some ignorance or reluctance on the part of the entanglee if 'entangling' with the relevant negative connotations is to be properly applied. Although we can imagine a case where Buyer is coerced by Seller into purchasing the latter's shares, the ordinary case of selling shares on the open market is quite different. For one thing, the mechanics of the electronic stock exchange put increasingly in doubt that shares have the kind of identity-conditions which secure the claim that any given Buyer is buying the very particular shares that some given Seller is selling. If there is a very small market in the shares, this may be possible; but if tens of thousands of shares are being regularly traded, it will not be possible. In the second place, Seller simply authorizes Broker to sell the shares; while Broker may know who Buyer is (though Broker may not, but sell only to another broker), Seller will not know. There are paradigms of harm being done to 'unassignable individuals', in Jeremy Bentham's splendid phrase — pollution of a river or of the air. The pulp mill may culpably know that some people will be harmed by its discharges although it does not know which people. If one is prepared to regard Seller's selling shares when it may not be possible to identify who buys them, or when, even if it is possible, Seller will be ignorant of Buyer, as contamination of 'unassignable individuals' comparable to a pulp mill discharging into a river, then I cannot confound that interpretation. But 'entangling' is already beginning to look like a tendentious description, and the distance of these cases from the paradigms of entangling is considerable. Third, as said, 'entangling' presupposes innocence or reluctance on the part of the entanglee. It is extremely unlikely that the ordinary case of shares being purchased on an open market involves any such ignorance or reluctance.

In short, cases where Seller really does 'entangle' Buyer and pass on the contamination of the soul, these are exceptional cases. What actually happens in the normal cases of a sale of shares would not support the description 'entangling' at all.

The second way of interpreting 'entangling' focusses on the harm done by the possession of the objectionable shares. Seller purportedly 'entangles' Buyer by causing Buyer to do exactly the same harm as Seller was doing. Indeed, as Gordon and Sadowsky note ([4], p. 155), the issue here is causation. Cahn's puzzle, and their support of it in this respect, require that, if Seller is to be rightly described as 'entangling', the causal connection from Seller to the harming investment remains unbroken by the act of selling the shares to Buyer. But it seems to me that this is simply not so. The voluntary purchase of the shares by Buyer on the open market is an intervening act which breaks the chain of causation from Seller to any harm done by the investment. Seller does *not* 'cause' Buyer to buy the shares, so long as Buyer has acted voluntarily in buying them.

The term 'cause' is slippery in this context. In one sense, it is perfectly true that Seller 'caused' Buyer's purchase (presuming the problem above of the identity-conditions for shares is solved). This is the sense of 'but for' causation, 'cause in fact', or necessary condition. 'But for' Seller's act of placing those shares on the market, Buyer could not have bought exactly those shares. Notoriously, however, the expression 'A caused B' means more than 'But for A's occurring B would not occur'. This sense of 'cause' is in common to both statements about inanimate ('The hail caused damage to the crop') and animate ('A's shriek caused B to jump') entities. The connection is termed by Hart 'causal responsibility', to distinguish it from 'legal liability-responsibility', the matter (crudely speaking) of a person's psychological relation to an occurrence (cf. [5], pp. 214–22). That the law, as well as ordinary talk, embodies this distinction between causal connections is not in doubt. The claim 'Seller entangled Buyer' requires more for its truth than simple 'but-for' causation. It requires genuine causal responsibility. But Buyer's voluntary intervention in the market negates Seller's causal responsibility.

Consider these two cases. (1) Vendor builds (it is indifferent whether the building is negligent, reckless, or intentional) a rickety porch on to a house. Vendor then sells the house directly to Purchaser, saying nothing when Purchaser fails to notice the condition of the porch. Shortly after, the porch collapses, injuring Lettercarrier. (2) Vendor builds the rickety porch as before, but commissions Realtor to sell the house. Purchaser purchases the house; Purchaser is merely a name on the contract to Vendor. Purchaser realizes the porch is rickety, but does nothing to fix it. Shortly after, the porch collapses, injuring Lettercarrier. In case (1), there is some plausibility to the claim that Vendor has 'entangled' Purchaser in the situation which led to harm to Lettercarrier. There is very little plausibility to this claim in case (2). Given Purchaser's voluntary purchase of the house, and awareness of the state of the porch, there is no longer any causal responsibility on the part of Vendor for the harm to Lettercarrier – even though it is true that, but for Vendor building the rickety porch, the harm to Lettercarrier would not have occurred. The case of Buyer's purchase of Seller's shares on the open market is analogous to case (2), not case (1).

In short, if 'entangling' is taken to refer to a causal connection between Seller and the harm Buyer is doing by investing in Armaments Manufacturer, then there is no such connection. Talk of 'entangling' is not justified.

Let us make one final attempt to reinstate the puzzle. We say to Seller, who we assume to have made a sincere speech along the lines of the above discussion, 'You still have not seen the point. You claim you have an obligation to divest yourself of your shares in Armaments Manufacturer. That is because you believe it is wrong to possess these shares because of the harm that this company's activities cause. But inevitably, if you *sell* your shares, then someone else will *buy* them. That person will then be doing the wrong you believe yourself now to be doing. So your action of selling is one that occasions a situation in which someone does wrong. It can never be a requirement of moral principle to occasion a situation in which someone does wrong. So you must be mistaken in thinking you have an obligation to divest yourself of your shares in Armaments Manufacturer.'

Seller can make the following reply. ' "Occasion" is a far cry from "entangle". I would indeed do wrong if I "entangled" another person in a situation in which that person did wrong. Whoever will own my shares will indeed do wrong. But that is not my problem. As things are, in no way do I force, coerce, entice, or otherwise influence Buyer, whoever Buyer may be, to buy my shares. In no way do I force, coerce, entice, or otherwise influence Buyer to do the wrong that Buyer will do by owning those shares. I "occasion" Buyer's obtaining the shares, if you like, if we suppose (improbably, in the face of the facts) that, but for my shares being on the market, Buyer would not invest in Armaments Manufacturer. It is Buyer's own free choice to purchase those shares and make that investment. Consider the case of Executioner, who comes to believe that capital punishment is wicked and so resigns. Executioner, as an astute reader of the political tea-leaves, knows full well the state will have no trouble hiring another executioner, and that executions will continue. But it is absurd to say that by that resignation Executioner is "entangling" the new appointee in the wrong of capital punishment. The very notion of "responsibility" which you employ to hold me responsible is one that implies Buyer's or the new executioner's free choice makes them, not I or Executioner, "responsible". Executioner and I treat the other as an end, not as a means, in acknowledging the freedom to choose. We do what in conscience has to be done. Executioner's conscience is now clear, and so is mine.' Seller might even have spoken of 'clean hands' as well.

We can now see that we are in familiar moral territory. Both Seller and Executioner are persons who place a high value on *integrity*. It is important to both Seller and Executioner that they lack certain kinds of connection to wrong things that go on in the world. It may even be more important to them that they are not connected with wrongs than that the wrongs do not occur. Cohen ([3], p. 110) says (self-righteously?) that a 'clean portfolio is merely a self-righteous [goal]'. But those who value integrity despise such cynicism. The cases of Seller and Executioner throw into sharp relief the conflict between integrity-based moralities and utility-based moralities. In the name of the latter, Shareholder will urge Seller not to sell, but to work from within to change investment and manufacturing policies. Reformer may urge Executioner to remain and use the authority of a public position to campaign for abolition. Of such issues are agonies and tragedies made, both real and fictional. But, as long as integrity is of moral value, then the believer in integrity can repudiate entirely successfully the charge of 'entangling' in divestiture cases of the kind at issue.

I have shown that morally principled divestiture is indeed possible — this will be divestiture which is not 'entangling', and I have shown how *that* is possible. The key principle is the principle of integrity. Cahn also ([2], p. 156) puts the task he sets as showing whether there is a 'duty' to divest, whether divestiture is required. To show how principled divestiture is possible is not the same task as that of showing that everyone has a duty to divest. That task would require for its completion a powerful argument of normative substance, and not the solution of a philosophical problem about the possibility of a coherent argument. Cahn's puzzle as first posed was the theoretical one, not the substantive one. I have shown how it is that someone who believes in the duty to divest may show that the duty may be done in a principled way. The divestiture puzzle as put forward is therefore dissolved.[1]

University of Alberta,
Edmonton, Alberta, Canada T6G 2E5

REFERENCES

[1] Steven M. Cahn, 'A Puzzle Concerning Divestiture', *Analysis* 47 (1987) 175–6.
[2] Steven M. Cahn, 'A Note on Divestiture', *Analysis* 49 (1989) 156–7.
[3] Daniel H. Cohen, 'A Reply to Cahn', *Analysis* 48 (1988) 109–10.
[4] David Gordon and James Sadowsky, 'Morally Principled Divestiture?', *Analysis* 49 (1989) 153–5.
[5] H. L. A. Hart, *Punishment and Responsibility* (Oxford: Clarendon Press, 1968).
[6] Kerry S. Walters, 'Morally Acceptable Divestiture', *Analysis* 48 (1988) 216–18.

---

[1] I would like to thank Don Dedrick and Janet Sisson for valuable comments and encouragement.

THE PRAGMATIC PROBLEM OF INDUCTION: REPLY TO GOWER AND BAMFORD

By JOHN WATKINS

I DO not understand why Barry Gower, in his 'Corroboration Versus Borrocoration' (ANALYSIS 49.1, January 1989), supposes that his borrocorationist principle constitutes any sort of difficulty for my proposed solution (ANALYSIS 48.1, January 1988) of the pragmatic problem of induction. A hypothesis is said by him to be the best borrocorated in its field if it is the negation of one that is the best corroborated. And his borrocorationist principle, which mirrors my corroborationist principle, says that, during the past, present and future of the human race, the overall success-rate of actions guided by well borrocorated hypotheses is *lower* than that of actions guided by hypotheses that are not well borrocorated. This principle is juxtaposed with a counter-borrocorationist principle which, like my counter-corroborationist principle, comes in two versions. In one version it says that, during the past, present and future (of the human race – this phrase will henceforth be omitted), the overall success-rate of actions guided by hypotheses that are *not* well borrocorated is *lower* than that of actions guided by well borrocorated hypotheses. In the other version it says that, in the future, the success-rate of actions guided by hypotheses that are *not* well borrocorated will be *lower* than that of actions guided by well borrocorated hypotheses. He introduces an historical assumption, call it *E*, which is analogous to one I introduced and which says that the success-rate in the past of actions guided by well borrocorated hypotheses has been *lower* than that of actions guided by hypotheses that are not well borrocorated. Gower now mimics my argument for there being a reason for a practical decision-maker to allow the borrocorationist principle rather than the counter-borrocorationist principle to guide his actions. He writes:

> Bearing in mind the historical assumption about the lower success-rate of well borrocorated hypotheses in the past, it can easily be shown that the counter-borrocorationist principle, in either of its versions, implies a stronger forecast about the future than does the borrocorationist principle …
>
> It follows, therefore, that a rational decision-maker confronting a practical problem should use the borrocorationist principle to guide actions (p. 9).

There are things in his presentation that I could quibble with (for instance, the mere negation of a well corroborated hypothesis won't offer the practical decision-maker any guidance); but assume that some reformulations would rectify matters without disturbing the main argument. Then Gower has shown that the pattern of reasoning that leads to the adoption of a corroborationist principle also leads to the adoption of a borrocorationist principle. What follows from this? It is well known that an inductive 'justification' for an inductive principle (or policy) can be mirrored by a counter-inductive 'justification' for a counter-inductive principle (or policy);[1] perhaps Gower thought that something similar has happened here. But it hasn't. We now have an argument for allowing a borrocorationist principle, as well as a corroborationist principle, to guide our actions; but *there is no conflict between the guidance they give.* Consider what they respectively say. Let H and $-H$ be, respectively, the best corroborated and the best borrocorated hypotheses in their field at the present time. Then the corroborationist principle says that decisions guided by H have a higher chance of success than decisions guided by $-H$; and the borrocorationist principle says that decisions guided by $-H$ have a lower chance of success than decisions guided by *H*. Perfect accord! If 'borrocorative reasoners' prefer successful to unsuccessful decisions, their decision-making will parallel that of 'corroborationist reasoners'.

Greg Bamford writes: 'CC_2 entails nothing about the past. So Watkins's belief that the strength of the forecast CC_2 makes about the future can be affected by the introduction of *E*, a proposition about the past, is an inductive belief' (ANALYSIS 49.4, October 1989). Here CC_2 says that in the future the success-rate of actions guided by hypotheses that are not well corroborated will be higher than that of actions guided by hypotheses that are well corroborated, and *E* says that in the past the success-rate of actions guided by hypotheses that are well corroborated was higher than that of actions guided by hypotheses that are not well corroborated. I inferred that CC_2 and *E* jointly imply 'that a time will come when success-rates change over and, moreover, that *this time has come now*'. This was not an inductive but a logical inference. (What Bamford says about the box with the red and blue balls and the uninspected compartments is true but beside the point. I did *not* say that CC_2 and *E* jointly imply a stronger claim about future success-rates than *C* and *E* jointly imply.)

Bamford asks: 'Why assume that it is better or more rational, other things being equal, to act on the weaker of two forecasts or assumptions about the future? Suppose I glimpse a snake in the grass ahead. Should I act on the assumption that the snake is venomous, or that it is venomous and aggressive?' I had said that it is rational to 'choose that course whose success presupposes the weaker forecast about the future'. Let the choice here be between (*D1*) giving the snake a wide berth, or (*D2*) walking straight on and passing fairly close to it. Assume that the snake is venomous, that it does not chase people who give it a wide berth, and that success consists in not getting bitten by it. Then the success of *D2* presupposes that the snake is not aggressive (walking fairly close to it won't provoke it to attack). The success of *D1*, by contrast, does not presuppose either that the snake is, or that it is not, aggressive. So the pragmatic principle on which I was relying recommends giving the snake a wide berth.

I am not sure that I have understood Bamford's last two paragraphs; he seemed to be rejecting a general pragmatic principle on the ground that it won't help him over his very next decision. That seems to be rather like complaining that a general advice to home-owners to insure their property doesn't tell me whether *I* should insure *this* house.

The London School of Economics,
Houghton Street, London WC2A 2AE

---

[1] For references see my *Science and Scepticism*, § 3.33.

Philosophy

Vol. 50 No. 4 (New Series No. 228) October 1990

ANALYSIS

Edited by
PETER SMITH

vidson and the sceptic: the thumbnail version EDWARD CRAIG
ammar, semantics and conditionals V. H. DUDMAN
An alleged analogy between numbers and propositions TIM CRANE
nsularity and the persistence of perceptual illusion PHILIP CAM
Reinterpreting images PETER SLEZAK
Externalism and experience GREGORY MCCULLOCH
He wants to try ALFRED R. MELE
Consequentialism and the unforeseeable future ALASTAIR NORCROSS
Utilitarianism and the 'punishment' of the innocent: the general problem SAUL SMILANSKY
Murphy and mercy JACOB ADLER
Nonbranching and nontransitivity DOUGLAS EHRING
On omniscience and a 'set of all truths': a reply to Bringsjord PATRICK GRIM

BASIL BLACKWELL · OXFORD AND CAMBRIDGE, MA

ISSN 0003-2638

ANALYSIS COMPETITION

Entries are invited for the first ANALYSIS Essay Competition. Papers may be on any topic broadly in keeping with the traditional concerns of the journal. An upper limit of 5000 words, including references and footnotes, will be strictly applied. Entries (three copies, prepared for blind refereeing, and clearly marked 'Competition') must be sent to the Editor to arrive no later than 1 May 1991. Authors should be aged 30 or under on that date*.

The first prize will be £500, with an additional £100 worth of books donated by Basil Blackwell. Further prizes totalling £500 will also be awarded, divided at the judges' discretion. The winning essays, with a selection of other entries, will be published together in the October 1991 issue of ANALYSIS.

*Entrants who believe that there are grounds for relaxing the age rule in their case may consult the Editor.

ANALYSIS 50.4 OCTOBER 1990

DAVIDSON AND THE SCEPTIC: THE THUMBNAIL VERSION

By EDWARD CRAIG

GRANT the whole argument. Grant that if any being interprets another's beliefs, it must do so in such a way as to make them in the main true, by the interpreter's own lights. Assume that there is an omniscient being who interprets us, and conclude that the beliefs it ascribes to us really are, in the main, true ones. Allow that the beliefs which such a being would ascribe to us are the beliefs which we really have. Conclude, finally, that our beliefs cannot involve massive error ([2], p. 201, [3], p. 317).

Where does all this granting, assuming and allowing get us in relation to scepticism? Nowhere at all—unless *we know what the beliefs are* that the omniscient interpreter ascribes to us. Otherwise all we know, everything being granted and allowed, is that the belief we evince when we assent to the sentence 'Here is a hand' is true, *whatever it may be*. And since that may, for all that has been shown, be something about the intentions of Cartesian demons, or computer-input to artificially sustained brains, or something else quite beyond our imagination, this "concession" should hardly worry a sceptic. If anything, it broadens the range of his scepticism, since it puts our knowledge of what it is that we believe at risk along with our knowledge of what is the case. (Earlier versions of scepticism did at least tend to have the former alone.)

But could it be that we do know what it is that we believe? Only, on Davidson's account, if we already know what is true, since that is the only way in which we could know what beliefs the omniscient being will ascribe to us. For us to be in that position, however, the sceptic must already have been defeated somehow or other, without recourse to Davidson's argument. Like Nozick's attempt (see [1]), that argument is therefore either impotent (if not slightly worse), or redundant.

What could induce someone, it may well be asked, to offer this argument as a *rebuttal* of scepticism? I can only conjecture: an over-concentration on a single sceptical formula, 'It could be that most of our beliefs are false', might lead one to think that the defeat of this formula defeats scepticism, however the victory is achieved. Far from it: establishing merely that our beliefs are mostly true, by a method which simply transmutes any doubt there may be about their truth into a doubt about what their content is, throws away the baby without even managing to get rid of the bathwater.

Churchill College,
Cambridge CB3 0DS

REFERENCES

[1] E. J. Craig, 'Nozick and the Sceptic: the Thumbnail Version', *Analysis* 49 (1989) 161–2.
[2] Donald Davidson, 'The Method of Truth in Metaphysics', in his *Inquiries into Truth and Interpretation* (Oxford: Oxford University Press, 1984), 199–214.
[3] Donald Davidson, 'A Coherence Theory of Knowledge and Truth', in *Truth and Interpretation*, edited by E. LePore (Oxford: Blackwell, 1986), 307–19.

GRAMMAR, SEMANTICS AND CONDITIONALS

By V. H. DUDMAN

GRAMMAR is a necessary preliminary to semantics. In the first place, any semantic theory is bound to presume *some* structure in the messages it analyses. Moreover, the success of the semantic theory is going to depend directly on getting this structure right, on recognizing the right components and slotting them together right. But the only structure a message has is the one imposed by language as the price of encodement as a sentence. And discovering this structure is the business of grammar. Furthermore, messages are exactly identifiable only as expressed in sentences. There is therefore no other way of bringing a semantic theory into direct confrontation with any kind of evidence than to require its analyses of messages always to explain the sentences used to express them. But explaining how messages determine sentences is the business of grammar. Willy nilly, then, the semanticist needs to be a grammarian.

A claim so large and simple needs dramatizing. Fortunately the homely topic of conditionals affords its spectacular illustration. Received semantic theories of conditionals all prove upon examination to presume a thoroughly unbelievable set of connections between sentences and their informational burdens (Section VI). But then, received semantic theories of conditionals all attribute the same structure to them, a ternary one, and this ternary theory is at considerable variance with grammatical fact (Section IV).

I speak bravely, but have to admit to having often urged this before, evidently always with persuasiveness that would be grossly overstated as minimal. True, some philosophers have favoured some of my *classifications*, notably in connection with 'if'. But appreciably fewer have been moved by my submission that their semantic theories presuppose false grammatical *analyses*.[1] My impression is that grammar is widely seen as a mere handmaid to semantics, at best affording helpful cues, hints, guides, clues or pointers as to the underlying semantic realities – whereupon it is a matter for judgement how much notice to take of its findings.[2] It is in hope of unseating this comfortable delusion that I now enter the lists again.

I

The 'if'-sentences (A) and (B) are both ambiguous between a habitual message, present or past, and a single-sequence-of-events message relating to the future:

(A) If the bough breaks the cradle will fall
(B) If the bough broke the cradle would fall.

This is because each is generated by two *encoding programs* which happen to overlap in their outputs. There is one program which generates these sentences from habitual messages while another, quite different, generates them from messages about the future. Now, grammar is the empirical science that investigates encoding programs.

Encoding programs have messages for inputs and sentences for outputs. When an encoding program generates a sentence *S* from a message *m*, I call *m* an *interpretation* of *S* and say that *m* is *encoded in S*. The connection, emphasized by Frege, between complete messages and complete sentences is that one sentence is generated from one message by one start-to-finish run through some encoding program.

It is in the nature of an encoding program that it will impose a certain structure on the informational burdens it processes. Indeed we can assign the content of a message to the speaker's originating intentions and its form to the rigours of encodement. And thus the grammarian, student of encoding programs, emerges as arbiter of the empirical question of a message's structure.

There is a lesson in (A) and (B): that English sentences do not determine their interpretations. It is a lesson yet unlearnt by those many semanticists who predicate their theories directly upon sentences.

---

[1] In [1], for example, Johnathan Bennett draws 'the line' where I draw one of mine, at any rate up to a point. But he happily continues recognizing 'antecedents' and 'consequents' both sides of this line, something I inveigh against in each of [2] to [6]. And his 'straight' and 'corner' notations perpetuate a ternary analysis which I maintain applies on neither side.

[2] 'I don't believe that grammar should be allowed such a firm hand on the steering wheel', Bennett wrote to me in response to my [7]. Whereas I see grammar in the driver's seat: when her hand isn't on the steering wheel, look out.

For demonstrating the craft of grammatical analysis I take as example the second of the two encoding programs mentioned above — the one which generates (A) and (B) from their future interpretations. First we need an impression of its dimensions. Review of the teeming evidence discovers it generating each of (C), (D) and (E) from single-future-sequence interpretations:

(C) If/provided that/whether or not/unless the bough breaks the cradle will/can/may/must/should$_1$/ought to/needn't/daren't fall
(D) If/provided that/whether or not/unless the bough broke the cradle would/could/might fall
(E) If/provided that/whether or not/unless the bough had broken the cradle would/could/might have fallen.

(We say (E) about the future when an intended ascent to the treetop with Baby has been called off.) This same encoding program is also responsible for generating (F) from a present interpretation, (G) from a present interpretation, and both (E) and (G) from past interpretations:

(F) If/provided that/whether or not/unless she was here, Her Majesty would/could/might be disgusted
(G) If/provided that/whether or not/unless she had been here, Her Majesty would/could/might have been disgusted.

Indeed it is easy to see the hand of a single encoding program in all this. (The terms 'indicative' and subjunctive', which would segregate the future interpretation of (C) from all the rest, find no justification in English grammar. See [6].)

When this program selects 'if' in particular, I call the message thus encoded a *conditional*, and throughout this essay I intend and have intended that travelled term in this exact and specialized sense. We see here another benefit of grammar, incidentally. It enables demarcation of the messages a semantic theory might apply to, in terms which are at once nonsemantic — hence independent — and exact.

No doubt the encoding program for conditionals has further accomplishments, unanticipated. But let us apply our grammatical skills to the data we have.

II

The first thing to see is that the encoding program for a conditional is merely an elaboration of a simpler one, namely that which generates (H), (I) and (J) from future-single-event interpretations, (K) and (L) from certain present-state interpretations, and (J) and (L) from certain past interpretations:

(H) The cradle will/can/may/must/should$_1$/ought to/needn't/daren't fall
(I) The cradle would/could/might fall
(J) The cradle would/could/might have fallen
(K) Her Majesty would/could/might be disgusted
(L) Her Majesty would/could/might have been disgusted.

Messages encoded by this program I call *projective.*

My proposal then is that conditionals are merely a proper subset of projective messages. Both are encoded by basically the same program. It is just that a conditional has an extra factor in it, calling for an extra episode in the encoding program. This episode inserts a subordinate clause into the sentence's predicate. Under their projective interpretations, (H) and (C) agree in outermost structure, both subject-predicate sentences. This analysis is confirmed by the observational fact that, under its conditional interpretation, (C) has the following stylistic variants:

The cradle, if the bough breaks, will/must/etc. fall
The cradle will/must/etc. fall if the bough breaks.

It will be observed that a projective message always requires a so-called *secondary auxiliary* for its encodement. Now, on general behavioural grounds the secondary auxiliaries can be arranged two-dimensionally as follows:

will	would
can	could
may	might
shall	should$_2$
must	—
should$_1$	—
ought	—
need	—
dare	—

It means that two separate informational factors are encoded into a secondary auxiliary: one to choose its row and the other its column.

III

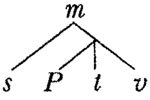
There is an easy part to grammatical analysis and a hard part. The easy part is discriminating informational factors in the message. The hard part is working out how they slot together. In this section I try to analyse the basic encoding program for projective messages, explaining a little as I go.

When (H) to (J) are being generated from their projective interpretations, we discover just four influences ringing the changes on them. Each particular variant sentence might have had a different subject, a different verb phrase, a different 'form' in the sense particularized in the next paragraph, or a secondary auxiliary from a different row, all quite independently; but there is no further avenue to variation in the message short of introducing more parts of speech into the sentence. I diagnose four informational factors in the message: s, responsible for the subject of the sentence, P, responsible for the verb phrase, t, responsible for the form, and v, responsible for the row of the auxiliary.

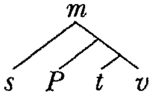
What I am calling the *form* is determined as follows. First, there is always the choice of a column for the secondary auxiliary. And if the right-hand column is chosen, there is the additional possibility of phase modification. *Phase modification* consists in replacing an inflectional form of a verb by the same inflectional form of 'HAVE' followed by the 'past participle' of that verb: 'broke' by 'had broken'; 'would fall' by 'would have fallen'.

So a projective message must have an s, a P, a t and a v, whatever the significance of those four cyphers. That much was easy, merely a matter of doling out informational factors wherever we recognised 'paradigmatic' alternatives. Now things get trickier.

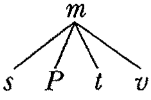
We have discriminated four factors, but are all *immediate* to the message? Perhaps — expectation encouraged by Frege's familiar concept-and-object analysis — the basic projective message has just two immediate factors, s and an amalgam of P, t and v (i). Or t and v might not even be immediate factors of an immediate factor of m, but only immediate factors of one (ii). And of course there are many other possibilities — among them the possibility that all four might be coequal (iii).



(i)



(ii)



(iii)

Arrangement (iii) is the staff solution. It accords with a basic maxim: make every factor as immediate to its message as possible, because the less branching it postulates the more flexible a theory is going to be. The onus probandi always lies with the proponent of the more elaborate analysis.

When there is an object to the verb, as in the projective interpretation of

(M) The cradle will break *the bough*,

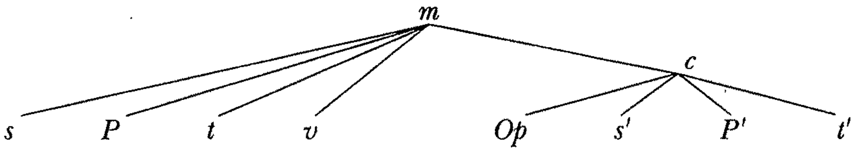
there has to be a factor — call it o — to determine it. Not immediate to the message, though. A new immediate factor is a more striking addition: think of introducing 'probably' into (H). Rather, o is immediate to P, along with the factor responsible for the choice of verb. In (M), in other words, the verb phrase is 'BREAK the bough'. We begin to appreciate the intricacy of the structural problem when we try identifying verb phrases in the wild.

IV

When the projective message is a conditional, the factors s, P, t and v are joined a fifth, the one encoded in the subordinate clause. I call it c. Naturally I postulate that c is immediate to the message.

We can dissect c if we like. It resolves into four factors: s', responsible for the subject of the clause, P' responsible for its verb phrase, t' responsible for its form, and Op responsible for the choice of 'if'. Presumably these four factors of c are all coequal.

In sum, *this* is the structure that I think English grammar discovers in the basic conditional:



V

With the grammatical structure in hand it is safe to proceed to semantics, and in this section I briefly instance my own semantic theory of conditionals as—whatever else—respectful of the structural realities. For explicitness, I apply it to examples.

The projective interpretations of (H), and (I) and (J) each have four immediate factors. The noun phrase 'the cradle' is chosen as the sentence's subject in order to identify the actual cradle as the *notional subject* s of the message. The verb phrase 'FALL' is chosen to help specify a *predication condition* P.[3] The form registers the *tense* t of the message, present past or past past. The secondary auxiliary's row is chosen to express a *verdict* v. And the message as a whole is a *judgement*, that is, a tensed verdict, concerning s's satisfying P at a time y that is later than t. (Notional subjects are Fregean objects. Predication conditions are conceptual affairs, and doubly unsaturated: satisfiable *by* a notional subject *at* a time. Verdicts are personal reactions of the speaker, which is why judgements are not candidates for truth.)

---

[3] Strictly, the predication condition is a conceptual elaboration $P(r)$ out of a *root condition* r, and the verb phrase is chosen to specify r: see [7] section 3. But I omit this complication as irrelevant to the present broad structural issue.

It has emerged that a message need not be about its tense. A projective message is intuitively about the time *y* of *P*'s satisfaction, and *t* is always earlier. This brings me to a central semantic proposal.

Among the messages English encodes in subject-predicate sentences, the fundamental division is into those which are intuitively about their tenses and those which are not. The former are *propositions*, the latter *judgements*. There is a basic structural difference: a judgement incorporates a verdict as well as the *s*, *P* and *t* a proposition has. Because of the personal nature of verdicts, judgements are not meant to be true or false. But a proposition propounds a historical fact about its tense, and hence can be transmitted truly or falsely. — These are enormous claims to be leaving undefended. But I wanted to disclose the full significance of my classifying conditionals as judgements.[4]

The traditional ascription of a 'future tense' to English consists, precisely, in conflating *t* with *y* in certain projective messages. It is grammatically untenable: if (N) below merely said about the future what (O) says about the past, there would be no accounting for the choice of *row*:

(N) The cradle will fall
(O) The cradle fell.

The traditional doctrine oversimplifies the projective interpretation of (N) by forgetting that 'will' in (N) is in paradigmatic relation to 'may', 'must' and so on.

Any theory of projective messages has to explain why *y* is later than *t*, since this trait is obviously essential to them. My proposal is that these judgements are verdicts about *s*'s satisfying *P* in a situation at *y* which is arrived at by starting from certain historical realities of *t* and imagining causal consequences to unfold futurewards from these realities. The speaker conducts a "fantasy" which sets out from facts of *t* and develops continuously to *y* in ways nature has taught her to expect. And then the verdict is reached by inspecting the contents of this fantasy.

When the projective message is a conditional, what is the subordinate clause up to? Inserted in the sentence's predicate, it encodes what I call a *complication* c of the otherwise simpler message. The subject of the subordinate clause is chosen to identify the notional subject s' of the complication. The verb phrase is chosen to help specify a predication condition P'. The form is chosen to register a tense t', present past or past past. And 'if' is chosen to announce an operator *Op*.

---

[4] In fact English has two basic programs for encoding judgements. We see the handiwork of the other in

Father will/can't/may/must/should$_1$/ought to/needn't
daren't/would/could/might be/have been asleep in his study.

Here *t* and *y* are each registered as either present or past quite independently, the former by the secondary auxiliary's column and the latter by the absence or presence of phase. The semantic difference is that these judgements are verdicts about *P*'s *actual*, historical satisfaction by *s*, whereas a projective message is a verdict about *P*'s satisfaction by *s* in a situation at *y* that can only be *imagined*. (See [4], sections 2, 30–33, 43–47.)

I submit that t' always locates the same historical point as t, even when—data unanticipated in Section I—this point is given as past in one place and past past in the other:

(P) If Her Majesty was here she would/could/might have been disgusted
(Q) If Her Majesty had been here she would/could/might be disgusted.

This is not impossible, because the past past *is* past.

The effect of the complication encoded with 'if' is to intrude into the fantasy the imagined satisfaction of P' by s' at a time x which is later than t' and hence later than t. The situation at y now has two strands in its causal antecedents: it results from s''s satisfaction of P' at x in circumstances developing out of realities of t. And then, as before, the judgement is a verdict about P's satisfaction by s in this situation at y.

—That, then, is how a semantic theory of conditionals might accommodate the grammatical structure I proposed for them at the end of Section IV.

VI

Diverse in other regards, received semantic theories of conditionals all agree in imputing the same ternary structure to them; diagnosing an 'antecedent', a 'consequent' and an operator. And logical notations all copy this same ternary analysis. In this concluding section I discover the ternary theorist committed unawares to a bewildering and patently arbitary complex of doctrines connecting conditionals and the English sentences that encode them.

Ternary theorists do not always particularize their 'antecedents' and 'consequents'. But it seems that in the case of the past-conditional interpretation of

(R) If the bough had broken the cradle would have fallen

they are the propositions, respectively, that the bough broke and that the cradle fell. (Thus, for example, David Lewis [8], pp. 2f.) Now why *these* propositions, with their simple past tenses? Plainly, because x and y are simply past. The antecedent has P' satisfied by s' at a tense which is identical with x, and the consequent has P satisfied by s at a tense identical with y.

Let us ask the grammatical question: How are antecedent, consequent and binary operator encoded in the 'if'-sentence? Plainly the prior propositions are responsible for the two subjects and the two verb phrases, and the operator—call it Op_1 —is responsible for 'if'. But for the rest?

A proposition's tense is essential to it, because propositions differing only in their tenses are different propositions. So if the consequent is to be encoded in the 'if'-sentence, then not only its *s* and its *P* but also its *t* must be encoded. That is to say, the tense of the consequent must somehow influence the formation of the 'if'-sentence. And the same, of course, for the tense of the antecedent. Now where in (R) might the influence of these two past tenses be felt? Plainly only in the forms. But also plainly, something else influences the forms as well, for if the forms directly registered the tenses we should have 'broke' and 'fell' in (R). This something else can only be Op_1. In short, Op_1 and the values of *x* and *y* jointly determine the form of (R)'s two predicates. (Thus Lewis loc. cit.) It might be that the relationship is more intimate, each proposition influencing only a proper substring ('the bough had broken' fixed by Op_1 and just the antecedent, 'the cradle would have fallen' by Op_1 and just the consequent). But a ternary theorist need not go so far. Suffice it to record the result that when *x* and *y* are both past, Op_1 requires 'had *U*-en' and 'would have *V*-en' for encoding the prior propositions. Not the least startling aspect of this result is the sudden accession of 'would'.

Turn now to the future-conditional interpretations of (A), (B) and (R). They have the same *s*, *P*, s' and P' and the same future *x* and future *y*, so they ought to have the same antecedent and the same consequent. Therefore we need different binary operators Op_2, Op_3 and Op_4 to account for the different 'if'-sentences. All three select 'if', of course. Then, with *x* and *y* both future, we have the results that Op_2 demands '*U*-s' and 'will *V*', Op_3 '*U*-ed' and 'would *V*', and Op_4 'had *U*-en' and 'would have *V*-en'.

It is a crucial question for the ternary theorist how Op_1 relates to our latest three. Which, if any, is it identical with, and how does it compare with the others? Such questions multiply when we consider Op_5 to Op_8, evidently required to account for the shared antecedent and consequent of the present-conditional interpretations of (P), (Q), (S) and (T):

(S) If Her Majesty was here she would be disgusted
(T) If Her Majesty had been here she would have been disgusted.

To say nothing of more recondite conditionals—with *x* present and *y* future, perhaps: various combinations occur.

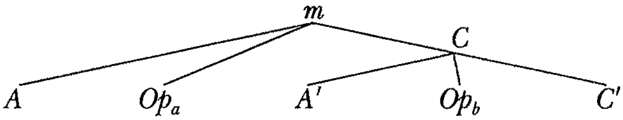
We are accumulating a veritable profusion of results connecting time and form. How many binary Op_is were needed to cover them all would depend on such things as how (P) and (Q) were handled.

As least three, for certainly none of Op_2, Op_3 and Op_4 could be identical. Presumably each of these Op_is would determine a pair of forms for certain — perhaps every — pair (x, y) of times. It would be interesting to see each of these Op_is delineated: supplied with both a meaning and a rationale for its assemblage of time-form arbitrations. And in truth the ternary theorist owes us nothing less. More, rather, for a deluge of new cases impends.

The ternary theory, which ascribes an intuitively past antecedent and consequent to the past-conditional interpretation of (R), simply does not provide for that conditional *itself* to be about a time of its own. But the ternary theory invokes iteration to explain examples like the past interpretation of

(U) If Grannie had removed the safety harness the cradle would have fallen if the bough had broken,

construing them as conditionals with conditional consequents:



Are Op_a and Op_b here both Op_1? In any case since Op_a's consequent C is itself a conditional, at least some of our labouring Op_is are even more accomplished than we thought, helping to assign forms *not* on the basis of time.

Evidently our Op_is analyse into two immediate factors, one shared by them all and another peculiar to each. The former factor is responsible for the common selection of 'if', and we need it to account for the indubitable intuition that 'if' means the same thing in all conditionals. It must be another, separate factor that manipulates the forms in the 'if'-sentence.

Then again there must be other binary operators which partake of the same range of second factors but have a different first one. We need a first factor for 'unless', another for 'whether or not', and so on. The suspicion arises that our binary operators are not authentic factors at all. Why should they be needed over and above their immediate factors? Indeed the basic maxim of Section III would have those two factors each immediate to the conditional, and the onus is on ternary theorists to explain why they reject a quaternary analysis.

One arresting corollary of received doctrine is that 'will' and 'would' are unrelated. For received doctrine has tense and 'mood' independent, indeed running athwart each other, as in 'present indicative', 'future subjunctive' and so on. But it is also received doctrine that 'will' in (A) registers 'future tense' while 'would' in (B) and (R) registers 'subjunctive mood'!

It is by assigning these special, quite fictitious roles to 'will' and 'would' that the ternary theorist avoids confronting them *as secon-*

dary auxiliaries. And this brings us to a really fundamental objection. Ternary theories confine themselves to only a narrow subclass of conditionals, rarely descending below even the top row of the table of secondary auxiliaries. Resisting the urge to generalize, they confront examples like (A), (B) and (R), rather than (C), (D) and (E). This is bad science: unignorable family resemblances connect 'will' with 'can', 'must' and the rest. And in the present case the narrow approach disguises a serious problem. Once it is seen that 'will' (say) is in paradigmatic relation with other secondary auxiliaries, it becomes hard to miss the need for a separate factor *v* to trigger the choice of row. And where might a ternary theory accommodate this *v*?[5]

Macquarie University,
Sydney, N.S.W., Australia 2109

REFERENCES

[1] Johnathan Bennett, 'Farewell to the Phlogiston Theory of Conditionals', *Mind* 97 (1988) 509–27.
[2] V. H. Dudman, 'Conditional Interpretations of "If"-sentences', *Australian Journal of Linguistics* 4 (1984) 143–204.
[3] V. H. Dudman, 'Parsing "If"-sentences', *Analysis* 44 (1984) 145–53.
[4] V. H. Dudman, 'Towards a Theory of Prediction for English', *Australian Journal of Linguistics* 5 (1985) 143–96.
[5] V. H. Dudman, 'Antecedents and Consequents', *Theoria* 52 (1986) 168–99.
[6] V. H. Dudman, 'Indicative and Subjunctive', *Analysis* 48 (1988) 113–22.
[7] V. H. Dudman, 'Vive la Révolution!', *Mind* 98 (1989) 591–603.
[8] David Lewis, *Counterfactuals* (Oxford: Basil Blackwell, 1973).

[5] I am indebted to the Editor for many improvements.

AN ALLEGED ANALOGY BETWEEN NUMBERS AND PROPOSITIONS

By TIM CRANE

A COMMONPLACE of recent philosophy of mind is that intentional states are relations between thinkers and propositions (see [4] p. 78; [5] p. 178; [6]; [3] pp. 7–9; [14] pp. 7–8; [13] p. 7). This thesis—call it the 'Relational Thesis'—does not depend on any specific theory of propositions. One can hold it whether one believes that propositions are Fregean Thoughts (see [7]), ordered n-tuples of objects and properties ('Russellian Propositions'; see [12]) or sets of possible worlds (see [9] and [14]). An assumption that all these theories of propositions share is that propositions are abstract objects, without location in space or time. So whichever theory of propositions is correct, the Relational Thesis says that intentional states are relations to certain abstract objects.

This thesis gives rise to one aspect of the problem of intentionality: how can intentional states, thus conceived, have any causal powers? Our commonsense theory of intentional states says that intentional states have effects that are appropriate to their propositional contents. But abstract objects have no causal powers. So how can the effects of intentional states depend on the abstract propositions they incorporate? Indeed, how can intentional states have any effects at all? This causal problem is one of the most intractable problems of intentionality (for discussion, see [4] p. 79; [5] pp. 201–23; [10] p. 60).

There are a number of options for a solution to this problem—for instance, one may deny the Relational Thesis, deny that intentional states are causes, or (the least promising) deny that relations to abstract objects cannot be causes. Or one may try and reconcile these three theses in some way—the approach I favour. But if one takes this approach, then one is obliged to show how a relation to a proposition can be causally efficacious. Or, less cryptically: one has to give an account of the causally efficacious properties whose instantiations make it the case that a thinker is related to a proposition.

Paul Churchland ([1] p. 105), Hartry Field ([4] pp. 113–4) and Robert Stalnaker ([14] p. 8) have argued that this apparent difficulty with the Relational Thesis may be disarmed by comparing ascriptions of intentional states to ascriptions of physical magnitudes like weights or temperatures. Just as 'X believes that *p*' may be thought of as expressing a relation—belief—between X and the proposition *p*, so 'X has a temperature of 100 degrees centigrade' may be thought of as expressing a relation—temperature in centigrade—between X and the number 100. According to this analogy, propositions are ways of 'indexing' beliefs and other intentional states, just as numbers are ways of indexing temperatures and other physical magnitudes (see [10] pp. 62–5 and [11] pp. 106–7 for this idea of indexing).

If this analogy is accepted, one could respond to it in one of two ways, the second more plausible than the first. One could say that there is nothing peculiar about invoking such real relations to abstract objects in intentional psychology, since they are invoked all the time in physics. Just as having a temperature is a real relation to a number, so having a belief is a real relation to a proposition. What makes it the case that an object has a temperature of 100 degrees is not just something about that object, but about its relation to the number 100. It is in virtue of standing in this relation that it has the temperature it does. And likewise with belief.

This cannot be right, because if it were, then the same problem would arise with temperatures (and weights etc.) that arises with intentional states. How could the state of something's having a certain temperature have effects, if it is really a relation to an abstract object? The causal problem of intentionality would seem to apply not only to intentional states, but to physical magnitudes as well! But of course, if any properties at all are causally efficacious, then temperatures, weights and the like are. So this way of interpreting the analogy is hopeless; it creates more problems than it solves. (See [8] pp. 17–18 and 253–5 for further arguments against regarding measurements of physical quantities as involving real relations to numbers.)

None of the above philosophers accepts this interpretation of the analogy. Stalnaker ([14] pp. 8–14) appears to do so, but in fact he is concerned with the relational structure of intentional and physical magnitude *predicates*; he reserves judgement on the 'ontological thesis about what kinds of properties are expressed by certain semantically complex expressions' (p. 10). Since this first interpretation of the analogy is incredible and has no serious defenders, I shall discuss it no further.

The more credible interpretation of the analogy, due to Churchland ([1] p. 105), attempts to undermine the Relational Thesis by pointing to the implausibility of the idea that having (e.g.) a temperature is being related to a number. So, the interpretation goes, there is no more reason to treat belief (etc.) as genuinely relational than there is to treat temperature (etc.) as genuinely relational.

This interpretation may be motivated by nominalism — the view that there are no abstract entities (see [4] appendix; [11] p. 141). The idea may be that the reason physical magnitudes are not relations to numbers is that there are no such numbers; similarly, the reason intentional states are not relations to propositions is that there are no such propositions. However, the interpretation does not entail nominalism, for one could believe in both abstract numbers and abstract propositions and still hold that the analogy shows that just as physical magnitudes are not really *relations* to these numbers, so intentional states are not really *relations* to these propositions. The issue is not about the existence of abstract objects, but about apparent *relations* between them and concrete objects. Nominalism, therefore, is irrelevant.

By suggesting that intentional states are not really relations at all, this second interpretation threatens to make short work of the causal problem of intentionality. It says that just as an object's temperature can be represented as its relation to a number, although it is patently not a relation, so a thinker's intentional state can be represented as a relation to a proposition, although it likewise is not a relation (see [10] p. 142). But the causal problem of intentionality was a consequence of the Relational Thesis; so if the analogy undermines this Thesis, then there may be less of a problem of intentionality than we initially thought. The apparently relational character of intentional states may perhaps be a fiction generated by a feature of the 'surface grammar' of their ascriptions — the fact that 'believes' and the other intentional verbs are 2-place predicates.

Now while I have no desire to make the problem of intentionality more difficult than it already is, I think that its causal aspect cannot be dissolved so easily. I do not deny here that a solution to the problem of intentionality may end up by dispensing — in some way — with relations to propositions. What I deny is that this can be done as simply as the analogists claim.

We should all agree that having (e.g.) a temperature is not a real relation to a number: it is a non-relational or intrinsic property. The reason we can represent it as a relation is that all temperatures belong to a family of intrinsic properties (see [14] p. 9). The members of this family — like that of other magnitudes such as *weight, mass*, or *velocity* — can be classified numerically, and to a certain extent their behaviour is mirrored by certain mathematical operations. (Why this is so is a hard question that I can ignore here.) Once we have decided on a unit of measurement, say centigrade, we may pick out each property in the family of temperatures with a two-place predicate '*x* has a temperature in centigrade of *n*' true of pairs of objects and numbers. But these are just ways we have of picking out the non-relational properties in the family — the fact that we can do this does not, on its own, show that there is a real relation of *temperature-in-centigrade* that an object has to a number. If the analogy with intentional states is to work, then *belief* (for example) must pick out a family of non-relational properties, just as *temperature* picks out a family of non-relational properties. As Stalnaker says,

> The analogy suggests that to define a relation between a person and a proposition is to define a class of properties with a structure that makes it possible to pick one of the properties out of the class by specifying a proposition. ([14], p. 11).

But does the analogy work?

In the case of physical magnitudes like temperature, which number indexes an object's temperature depends on the choice of a unit. But there are other actual systems of measurement of temperature, under which the object's temperature is indexed by different numbers. Which numbers are the right ones? All of them? Or is there some privileged number?

It is implausible to say that any one number is privileged. Take weight: a standard bag of sugar weighs 2.2 pounds or 1 kilo. Nationalistic prejudices aside, statements of weight in imperial and metric units are both equally statements of *weight* — and both equally true, when they are true. So the bag may be described as being related to 2.2 by the relation 'weight-in-pounds'; and to 1 by the relation 'weight-in-kilos'. Which number the bag is 'related to' depends on which unit of measurement is in question.

What is the analogue of this in the case of intentional states? It is true that which proposition I am related to depends on which *attitude* is in question — belief or desire or whatever. For instance, I may be related to the proposition *There is no beer* by the relation *belief*, and to *I have some beer* by the relation *desire*. But of course, this belief and this desire are different states, whereas there is an obvious sense in which '2.2 lbs' and '1 kilo' are different ways of picking out the *same* state of the bag of sugar — its weight.

The analogue of a unit of measurement in the case of intentional states is rather this. Anglophone Alf's belief that snow is white may be described as a relation between Alf and an abstract object, the English sentence-type 'Snow is white', tokens of which Alf is disposed to utter from time to time. But a French speaker can classify his belief by relating it to a different abstract object, the French sentence-type 'La neige est blanche'. These are two ways of indexing Alf's one belief, and (like farenheit and centigrade, or pounds and kilos) there can be translations between these different indexings. So like weights, intentional states can be classified by being related to abstract sentences. And like the numbers that measure weights, no one sentence is privileged (nationalistic prejudices aside) and the same intentional state can be indexed by any number of different sentences, depending on the choice of a language (the 'unit of measurement').

Those who are sceptical of propositions may use this analogy between sentences and numbers to illustrate how intentional states can *only* be indexed by sentences, not propositions, and thus only relative to a public language. (Compare Davidson's use of the analogy: [2] pp. 224–5.) But this claim does not, of course, follow from the mere analogy alone, and arguments against the existence of propositions that may support it are largely irrelevant to the problem of this paper. After all, the problem is how a relation to an abstract object can be a cause; and this problem arises, on the face of it, with sentence-types as with propositions alike. The issue, as I said above, is about the causal efficacy of relations to abstract objects, not about the existence of those objects.

So just as numbers can index temperatures relative to a unit of measurement, so sentences can index intentional states relative to a language. But propositions are not sentences. They are supposed to be what sentences with the same meanings express; and they are supposed to be the contents of intentional states of all thinkers, whether or not they speak a public language. So propositions should not be thought of as indexing intentional states relative to a language.

This fact reveals exactly where the analogy between numbers and propositions breaks down: with propositions themselves, unlike the sentences that express them, there is no analogue of the unit of measurement. It makes no sense to say that a given belief, say, can be related to many different abstract propositions, as it can be related to different abstract sentences, since if a belief were related to a different proposition, it would be a different belief. *Pace* Davison, the natural language sentences that index beliefs are no more essential to those beliefs than the particular numbers that index token weights and temperatures are essential to them. But the propositions that index beliefs are thus essential. The proposition to which a given belief is related is part of what individuates it, part of what makes it the belief it is. This is because the proposition gives the conditions under which the belief is true, and beliefs are standardly individuated by their truth conditions. Desires, likewise, are individuated by propositions that give their satisfaction conditions. (Some may deny these truisms, but they will need more than the analogy to prove the point.)

The disanalogy, then, is this. All that is essential to the 'relationality' of temperature is that each token temperature is index*able* by some number or other. Which number this is is relative to the unit of measurement in question. But token intentional states are not merely indexable by some proposition or other: each token is essentially index*ed*, without relativisation, to a *particular* proposition.

But why does this disanalogy matter? Because it reveals what does seem to be the genuinely relational character of intentional states, as opposed to the pseudo-relational character of physical magnitudes. A thing's temperature may be indexed by some number or other, which number depending on the choice of a unit: it is *only* the arbitrary choice of a unit that enables temperatures to be indexed by numbers at all. But real relations do not have this 'unit-relativity'. Take *being a son*: each son is related to one particular father. But it makes no sense to say that token sons are 'indexed' by some father or other, which one depending on (some analogue of) the choice of a unit. Of course not: there is no analogue of a unit here. It is because being a son is a genuine relational property, while having a temperature isn't, that unit-relativity is essential to the pseudo-relationality of temperature, and inapplicable to the real relationality of being a son.

(Do not be misled by the fact that spatial relations can be measured — i.e. that some real relations as well as non-relational properties can be indexed by numbers. The two-place real spatial relation that holds between New York and London is not unit-relative; the three-place pseudo relation between New York, London and a number is. In any case, the relevant analogy is between the abstract numbers that index *non-relational* properties and the abstract propositions that index intentional states. After all, it would not help the solution of the causal problem of intentionality to be shown that intentional states were like *spatial* relations!)

What we have discovered is that what applies to sons here applies to beliefs. It makes no sense to say that token beliefs are just indexed by some proposition or other, which one depending on (some analogue of) the choice of a unit. For there is no analogue of a unit here either. Like sons, beliefs and desires do not have unit-relativity. And this structural similarity between beliefs and other relations strongly suggests that the relationality of beliefs is more than superficial.

So if the propositions that index token intentional states are essential to them in a way that the numbers that index token weights and temperatures are not, then the relationality of intentional states cannot be dismantled along the same lines as the apparent relationality of physical magnitudes. This is the crucial disanalogy between intentional states and physical magnitudes, and it frustrates the attempted dissolution of the causal problem of intentionality. So the analogy between numbers and propositions does not undermine the Relational Thesis; the causal problem of intentionality remains, unfortunately, as it was.[1]

University College London,
Gower Street, London WC1E 6BT

REFERENCES

[1] P. M. Churchland, *Scientific Realism and the Plasticity of Mind* (Cambridge: Cambridge University Press, 1979).
[2] D. Davidson, 'Reality without Reference', in his *Inquiries into Truth and Interpretation* (Oxford: Oxford University Press, 1984).
[3] D. Dennett, 'Beyond Belief', in *Thought and Object*, edited by A. Woodfield (Oxford: Oxford University Press, 1982).
[4] H. Field, 'Mental Representation', reprinted, with an appendix, in *Readings in the Philosophy of Psychology* vol. II, edited by N. Block (London: Methuen, 1981).
[5] J. Fodor, *Representations* (Hassocks: Harvester Press, 1981).
[6] J. Fodor, 'Fodor's Guide to Mental Representation', *Mind* 94 (1985) 76–100.
[7] G. Frege, 'The Thought: a Logical Inquiry', trans. A. & M. Quinton, *Mind* 65 (1956) 289–311.
[8] H. Kyburg, *Theory and Measurement* (Cambridge: Cambridge University Press, 1984).
[9] D. Lewis, 'Attitudes *de dicto* and *de se*', *Philosophical Review* 88 (1979) 513–43.
[10] B. Loar, *Mind and Meaning* (Cambridge: University Press, 1981).
[11] C. McGinn, *Mental Content* (Oxford: Blackwell, 1989).
[12] N. Salmon, *Frege's Puzzle* (Cambridge Mass.: MIT Press, 1986).
[13] S. Schiffer, *Remnants of Meaning* (Cambridge, Mass.: MIT Press 1987).
[14] R. Stalnaker, *Inquiry* (Cambridge, Mass.: MIT Press 1984).

---

[1] I am grateful to Joseph Melia, Hugh Mellor, Peter Smith and Jamie Whyte for very helpful discussion.

INSULARITY AND THE PERSISTENCE OF PERCEPTUAL ILLUSION

By PHILIP CAM

MANY perceptual illusions persist in the face of conflicting knowledge or belief. This is a fact often commented upon particularly in discussions of visual illusion.[1] The Muller-Lyer illusion, for instance, persists in the face of a well-confirmed belief that the two lines are of equal length; figures in the Ames room fail to maintain size constancy and the room tends to go on looking rectangular even though we know that the figures are normal and that the room is not rectangular; a hollow mask, obviously hollow upon close inspection, will persist in looking like a normal protruding face at a distance; a three-dimensional simulation of the Penrose triangle goes on looking paradoxical even given knowledge of its physical structure, and so on. In discussing his claim that perceptions are hypotheses, Richard Gregory notes this feature of perceptual illusion and suggests the beginnings of an explanation:

> Evidently we adopt certain assumptions in order to make use of available sensory information, for 'constructing' a perception. In situations for which these assumptions are not justified, the perception can be not only wrong but even logically impossible. This can occur even when there is a simple 'solution'; and even when the 'solution' is known intellectually or by other sensory information. This shows that if perceptions are hypotheses, they are not always the same hypotheses as those we accept intellectually. Also, intellectual knowledge may not correct perception. ([4], p. 205)

Gregory's suggestion seems to be that perceptions are 'constructed' by reference to assumptions that may persist, and continue to govern perception, even when the ensuing perceptual hypotheses fail to accord with intellectual knowledge. The persistence of perceptual illusion, then, results from the failure of intellectual knowledge to correct perceptions based on erroneous assumptions.

Even if Gregory's remarks point us in the right direction, however, they leave important questions unanswered. In particular, why are the assumptions employed by perception not overriden by so-called intellectual knowledge? This is the question I want to consider, and I will begin by stating three possible lines of reply to it. While these three suggestions may not cover all the possibilities, they do raise what seem to be the central competing intuitions as to why perception is somewhat impervious to intellectual knowledge. First, the answer could lie in the fairly straightforward matter of a lack of physical connections. If the systems responsible for perceptual analysis and those that store and manipulate intellectual knowledge are at least somewhat physically discrete, then perceptual analysis could avail itself of intellectual knowledge only if there were appropriate physical connections to serve as a conduit for communication. And these connections may not always exist. Nor would this state of affairs be altogether unexpected. If normally reliable perception can be effected by limited downward processing, that would certainly be conducive to the rapid, efficient operation of perceptual systems. Secondly, it might be that the relevant information is normally available to perceptual analysis, but that the system operates in such a way that this information is sometimes overriden. For instance, the relevant assumptions might tend to prevail over intellectual knowledge only when they are highly conditioned. Finally, as Gregory himself goes on to suggest, perhaps the explanation is that 'human language has so enriched intellectual understanding, as to put it beyond the reach of the essentially primitive, prehuman, perceptual hypotheses'. ([4], p. 210) The basic idea here is that the perceptual processing systems have a relatively primitive competence and at least sometimes cannot readily assimilate those outgrowths of cognition that language affords.

---

[1] For discussion of the examples that follow, see [4] and also Gregory's [3] and [5].

While these suggestions are by no means equivalent, the first and third of them do share an important feature in common, which is lacking in the second. The feature that I have in mind has been highlighted by Jerry Fodor in his discussion of the *informational encapsulation* of systems mediating what he calls 'input analysis'.[2] According to Fodor, the human cognitive apparatus, broadly construed, can be divided into three basic types of systems. These are (1) *transducers* which map proximal stimuli onto covarying neural firing patterns; (2) *input systems* which respond to transducer outputs and perform at least partial perceptual analyses; and (3) *central systems* which accept the outputs of input systems and mediate higher cognition or thought. The input systems, Fodor tells us, have a range of characteristics in common which they do not share with central systems. The chief amongst these is that input systems are informationally encapsulated, which is to say that they have rather restricted computational access to information that is available for computation elsewhere in the cognitive system, and in particular in higher cognition.

Now in regard to the persistence of perceptual illusion, an input system might be informationally encapsulated either because it lacks appropriate efficient connections to central systems that deal with intellectual knowledge (the first suggestion) or else because it has limited competence in utilizing the information that those systems provide (the third suggestion). Either way, the impact of intellectual knowledge upon the analysis of sensory material will be limited because of limitations of access. Yet while the notion of informational encapsulation is no doubt useful for cognitive theorizing, I wonder whether Fodor is right in thinking that the persistence of perceptual illusion provides examples of it. For a start, we would do well to question Fodor when he says that, in accounting for the persistence of perceptual illusion, 'it is hard to see an alternative to the view that at least some of the background information at the subject's disposal is inaccessible to at least some of his perceptual mechanisms'. ([1], p. 66) For informational encapsulation covers only two of the three possibilities mentioned above, and the other option is not without grounds to recommend it. For instance, it is worth noting that the background assumptions upon which thse perceptual illusions appear to depend are arguably all highly conditioned by experience. Rooms are nearly always rectangular (at least in our culture), faces protrude, solid shafts that appear to join do join, and so forth. Stimulus configurations whose retinal projections conform to these requirements, and exhibit no obviously counteracting cues, might succumb to these assumptions if the perceptual system had even a marginal preference for homebased assumptions over imported information. A homebased preference might do just as well as ignorance of intellectual knowledge to account for the effects.

---

[2] See [1], especially pages 64–86. For the persistence of perceptual illusion as an example of informational encapsulation, see page 66. See also Fodor's [2].

Ambiguous stimuli provide a potential source of insight here. Stimuli such as the Necker cube and the well-known duck-rabbit present the visual perceptual system with alternative possibilities between which it tends to have no overriding preference. Presumably that is why these configurations so easily elicit the so-called Gestalt switch between interpretations. Now it is in this kind of circumstance, where the perceptual system has no overriding preference between interpretations, that the system will more readily take instruction from outside. Often with very little effort one can intellectually manipulate perception from one interpretation to the other. Roughly speaking, though, the greater the perceptual preference for an interpretation, the more recalcitrant to instruction the perceptual system becomes. (One finds it hard to engineer the switch, or to hold onto the less favoured interpretation once it is attained.) Now these characteristics of ambiguous figures suggest that there may be no discontinuity between the two sorts of cases that we have been considering. Rather than having informational encapsulation in some cases, and permeability in others, and the difficulty attendant upon sorting out why that might be, perhaps we should posit a single principle to cover them all. The principle would be (roughly) that perceptual analysis will accept instruction to the extent that its own preferences are not clear and univocal. Indeed it seems to me that because the appeal to preference weightings might plausibly account for the full range of cases, and is to be countenanced in any event, it is a fair working hypothesis and should probably be favoured over informational encapsulation in accounting for the persistence of perceptual illusion.

If we accept this it suggests a basic characteristic of the mechanisms involved, a sort of narrow-mindedness or *insularity*. The first call of perceptual systems is upon their own resources, and they tend to be guided by information or suggestions intruding from other 'authorities' only to the extent that decisions based upon their local information base leave open further or alternative possibilities.

Now that we are focusing upon insularity, notice that the persistence of perceptual illusion after the intellectual discovery of illusion doesn't, in turn, undermine belief. In the case of the Muller-Lyer illusion, for instance, we persist in believing that the two lines are of equal length regardless of the fact that the information intruding from visual perception goes on favouring a different belief. So it seems that it is not only perceptual systems which exhibit insularity. The system incorporating our intellectual belief is also decidedly insular.

Insularity, then, is a quite different property from Fodor's informational encapsulation. For Fodor, the persistence of perceptual illusion is evidence of the informational encapsulation of input systems, but not, of course, evidence for the encapsulation of central systems. Central systems, on his view, have unrestricted access to the outputs of input systems. By contrast, as I have just been suggesting, the persistence of both perceptual illusion and contrary belief argues for the mutual insularity of both types of system.

My suggestion, then, is that the central cases of persistent perceptual illusion, thought to reveal informational encapsulation, are more plausibly treated as arising from insularity. This interpretation, I claim, has the twin benefits of (1) providing a single principle to cover both the persistence of perceptual illusion and the malleability of perception in cases of figural ambiguity, and (2) providing an explanation not only of the persistence of illusion in the face of contrary fixed belief but also of the fixity of belief in the face of contrary persistent illusion.

Genuine informational encapsulation no doubt occurs naturally in the brain. My only argument with encapsulation here is that it is doubtful whether the persistence of perceptual illusion supplies examples of it. On the positive side, I have attempted to suggest that a rather different characteristic, that of *insularity*, may be of some importance in coming to understand the modularity of mind.

University of New South Wales,
Kensington, Australia 2033

REFERENCES

[1] Jerry A. Fodor, *The Modularity of Mind* (Cambridge, Mass.: MIT Press, 1983).
[2] Jerry A. Fodor, 'Observation Reconsidered', *Philosophy of Science* 51 (1984) 23–43.
[3] Richard Gregory, *The Intelligent Eye* (London: Weidenfeld and Nicolson, 1971).
[4] Richard Gregory, 'Perceptions as Hypotheses', in *Philosophy of Psychology*, edited by S. C. Brown (London: Macmillan, 1974), 195–210.
[5] Richard Gregory, *Eye and Brain* (London: Weidenfeld and Nicolson, 1977).

REINTERPRETING IMAGES

By PETER SLEZAK

I INTRODUCTION

ANDY CLARK has recently [4] cited experimental work in psychology by Chambers and Reisberg [2] in support of his earlier arguments [3] for the view that 'the image is bound to an interpretation in a way that the percept is not'. That is, imagery is not quasi-perceptual, nor does it require some distinct pictorial medium of representation. Chambers and Reisberg found that subjects who had never seen ambiguous figures such as the duck-rabbit were unable to reverse their mental image of it. Nevertheless, the same subjects *were* able to notice the alternative interpretation in their own drawing of their image. These results of Chambers and Reisberg are significant in posing a challenge to the orthodox pictorial account of the representational medium subserving imagery (Kosslyn et al. [14]). Specifically, the Chambers and Reisberg results clearly suggest that images are not uninterpreted surrogate 'objects' which require perceptual processing and interpretation as in veridical seeing. Instead, images are intrinsically bound to an interpretation and it is in virtue of this interpretation that Clark says 'there is more to images than meets the eye' ([4] p. 56). Though Clark does not expand on this sense in which there is more to images, the idea is familiar from Pylyshyn's writing on the subject and his 'tacit knowledge' alternative account of imagery. Indeed, Pylyshyn [17] had precisely anticipated such results as those of Chambers and Reisberg ten years ago where he wrote:

> Although there are similarities between cognitive operations in perception and in imaging, there are also some outstanding differences that may be more revealing of the underlying processes. For example, the order of scanning and the sorts of things that can be "noticed" in imaging are much more constrained than in perception. The reason for this is partly that a scene has a stable independent existence and can be re-examined at will to produce new interpretations. In

> contrast, the construction of an internal description from stored knowledge can hardly be divorced from its interpretation. While some reinterpretation is certainly possible, it surely is more like the derivation of new entailments from the stored knowledge than like the discovery of new aspects of an environment by the usual visual means. Discovering even moderately novel readings from a mental image...has been shown to be exceedingly difficult. ([17] p. 37)

For Pylyshyn, images are something which must be more abstract, more conceptual and more highly cognitive than has been allowed on the orthodox pictorial, quasi-perceptual account. In the foregoing remarks, Pylyshyn anticipates the issue which has recently become a focus for the debate about the 'format' of images, – that is, the issue concerning the possibility of making novel readings from mental images. Accordingly, in view of this recent debate and its notorious antecedents, the matter is not settled quite as simply or straightforwardly as suggested by Clark's mere citing of the Chambers and Reisberg results. While I wish to support Clark's position 'on the side of the angels', it is important to acknowledge the competing experimental evidence and, further, to give some theoretical analysis which can reconcile the apparently conflicting data. Thus, on the positive side for pictorialism, Pinker and Finke [16] have reported subjects' ability to 'see' novel properties which should emerge in images only after they had been mentally rotated. Subjects were asked to perform tasks which demanded 'the ability to "see" the two-dimensional patterns that should emerge in their images if the images depicted the new perspective view accurately' ([16], p. 244). Finke and Slayton [19] extend this work, providing further evidence 'that people are capable of making unexpected discoveries in imagery' and that novel patterns can "emerge" from within imaged patterns. Most recently, Finke, Pinker and Farah [8] have sought to reinforce these claims that images can be inspected and reinterpreted by subjects 'applying shape classification procedures to the information in imagery' ([8], p. 51). This latter research has been specifically designed to counter the skeptical claims of Chambers and Reisberg [2] cited by Clark. Finke, Pinker and Farah acknowledge that 'the controversy over imagery has largely concerned whether images are like early perceptual representations containing information about the geometric properties of visual inputs, or like later cognitive representations containing information about the conceptual categories of interpreted objects' ([8], pp. 52–3). In supporting the first of these alternatives, they are concerned with defending the possibility of 'inspecting' images and thereby 'reading off' properties from the pictorial display in a close analogy with perception.

In this way, the question of re-interpreting images has become a new testing ground for the 'imagery debate' since the possibility of doing so follows as a direct implication of the pictorial view of the medium: On this view, the image is like the perceptual representations of early vision in the sense of being a raw display in a visual buffer having spatial, geometric properties and, therefore, in need of interpretation in the first place. Kosslyn explains that his model entails that 'images depict visual information, and that this information is interpreted by some of the same sorts of classificatory procedures used in classifying sensory input during vision' ([10], p. 52). Clearly, a conception of images as a quasi-pictorial 'surface representation' which requires that they must be interpreted in the first place, will be likely to permit their *re*-interpretation.

Undeniably, an extensive body of empirical research over two decades has appeared to support the pictorial, quasi-perceptual account against the largely 'philosophical' arguments of Pylyshyn and others with whom Clark has sided. In view of the apparent weight of this empirical case, whatever the merits of the contrary philosophical arguments, it would be helpful if this alternative account could be supported with experimental results of its own. Accordingly, I will indicate preliminary results of work which I have been conducting which provides further compelling evidence in support of the Chambers and Reisberg results and the descriptive or conceptual, tacit knowledge position. Specifically this work casts serious doubt on the postulate of a pictorial medium and the claimed abilities of subjects to rotate, inspect and re-interpret their images.

II PERCEPTS, OBJECTS AND HOMUNCULI

Before looking at these experiments, we may attend briefly to some of the background theoretical, conceptual and methodological problems which have plagued the debate. We see a telling symptom of the difficulties inherent in the pictorial medium approach where Finke persistently speaks of 'mental images as being functionally equivalent to physical objects or events, with respect to certain types of effects' ([6], p. 113). However, taken *literally*, this is nonsense since Finke is unlikely to be seriously concerned with comparing the properties of mental representations *qua psychological entities* with those of tables and chairs. This systematic category mistake in Finke's discussion reflects a significant theoretical commitment to taking images as some kind of direct, literal object of apprehension. Philosophers at least will be familiar with the difficulties inherent in such notions from theories of sense-data which were taken to be the direct, immediate objects of perception. Indeed, the so-called 'argument from illusion' is, properly speaking, an argument from hallucination — that is, *images*.

Kosslyn ([13], p. 149) explains one purpose of imagery involves 'recognition processes' to discover information which is not stored explicitly in memory and thus we 'look' at our images in a way which is analogous to the way we look at external objects in order to inspect them. Notoriously, such perception-analog accounts are prone to the homunculus problem, and repeated confident assertions that this danger has been avoided in recent accounts cannot be taken at face value. Mere disavowals and an awareness of the problem *in principle* are not enough to avoid the dangers *in practice*. Thus, Pinker [15] in his survey of theories of visual cognition asks how the mind could contain images without some little man in the head to look at them and he asserts:

> This is simply not a problem under the computational theory of mind: images may be construed as data structures, and there is no more of a conceptual problem with positing mechanistic operations that can access those data structures than there is in positing mechanistic operations that access other mental representations such as linguistic or logical structures, or positing operations that access data in a computer. ([15] p. 38).

However, in the first place, adverting to these other domains in which data structures need to be accessed does not help here because they are equally prone to the explanatory dangers of the homunculus problem. Secondly, their mere use of mechanistic computational notions is not ipso facto sufficient to avoid the homunculus problem in practice; these approaches have only showed us how they *might* do so, and whether or not the explanatory vacuity is *actually* avoided depends on the specific character of the models postulated. Being realized in a computer program would only make a theory ipso facto immune to the homunculus problem in the case when the model was a more or less complete working simulation of the cognitive phenomenon in question. The computational idiom in which current imagery theories are articulated remains far from this ideal.

III RESEMBLANCE

It is through the postulate of a picture-like representation which refers by means of *resembling* its referents that the traditional difficulty arises. The explanatory problem is familiar from Chomsky's related complaint against traditional grammars which implicitly relied on the linguistic knowledge of the *user or theorist* rather than explicit theory. Here too, in a similar fashion, the conception of a pictorial medium which refers through resemblance inherently requires the contribution of the theorist to provide the illusion that the theory works. This is because the notion of resemblance can do no work in the theory on its own as a purported mechanism of reference; mental representations cannot intrinsically resemble their objects since the resemblance of one thing for another requires that a *judgement* or comparison can be made by a third party external to them both. However it is just this comparison which is impossible in the case of mental images, since we have no independent vantage point from which to judge our own images in order to make the comparison.

It would seem that the theoretical postulate of resemblance is an illegitimate extrapolation from a misguided introspection. We can see the mistake clearly in Finke's equivocations: Thus, he begins his recent book acknowledging 'that imagery is a subjective phenomenon' ([7], p. 1) and appropriately defines mental imagery as 'recreation of an experience that in at least some respects resembles the experience of actually perceiving an object or an event' ([7], p. 2). So far so good: Notice here that Finke has only asserted the resemblance of one kind of *experience* with another. However, Finke proceeds to endorse a piece of dubious, if compelling, folk psychology according to which most people claim that their images often resemble *actual objects*. Thus, he concludes that 'a mental image must be more like a "picture" of something than a verbal description' ([7], p. 4).[1] However, this is simply a non-sequitur since Finke is only entitled to assert that a mental image is more like the *experience of seeing* a picture.

The problem concerning the resemblance between images and objects is familiar to philosophers from Berkeley's answer to Locke:

> But say you, though the ideas themselves do not exist without [i.e. outside] the mind, yet there may be things like them, whereof they are copies or resemblances...I answer, an idea can be like nothing but an idea; a colour or figure can be like nothing but another colour or figure. If we look ever so little in our own thoughts, we shall find it impossible for us to conceive a likeness except only between our ideas. ([1], p. 64).

Berkeley was not the only philosopher long ago to have warned against the very difficulties which beset recent psychological theories. Striking in its contemporary pertinence and, therefore, worth quoting at length is a less well known passage from another philosopher who appears to speak directly to Kosslyn and Finke et al. from the past:

> ...it is necessary to beware of assuming that in order to sense, the mind needs to perceive certain images transmitted by the objects to the brain, as our philosophers commonly suppose; or, at least, the nature of these images must be conceived quite otherwise than as they do. For, inasmuch as [these philosophers] do not consider anything about these images except that they must resemble the objects

---

[1] In passing it should be noted that the tacit knowledge and descriptive accounts of imagery have not been committed to the exceedingly implausible view that the descriptions in question are 'verbal' in any sense deriving from natural language. This is a straw man which is consistently presented as the opposing account by pictorialists.

> they represent, it is impossible for them to show us how they can be formed by these objects, received by the external sense organs, and transmitted by the nerves to the brain. And they have had no other reason for positing them except that, observing that a picture can easily stimulate our minds to conceive the object painted there, it seemed to them that in the same way, the mind should be stimulated by little pictures which form in our head to conceive of those objects that touch our senses; instead we should consider that there are many other things besides pictures which can stimulate our thought, such as, for example, signs and words, which do not in any way resemble the things which they signify. And if, in order to depart as little as possible from currently accepted beliefs, we prefer to avow that the objects which we perceive must truly transmit their images to the inside of our brain, we must at least observe that there are no images that must resemble in every respect the objects they represent—for otherwise there would be no distinction between the object and its image—but that it is sufficient for them to resemble the objects in but a few ways, and even that their perfection frequently depends on their not resembling them as much as they might. ...So that often, in order to be more perfect as images and to represent an object better, they must not resemble it. Now we must think in the same way about the images that are formed in our brain, and we must note that it is only a question of knowing how they can enable the mind to perceive all the diverse qualities of the objects to which they refer; not of [knowing] how the images themselves resemble their objects; ... ([5], pp. 89–90).
>
> Thus you can clearly see that in order to perceive, the mind need not contemplate any images resembling the things that it senses. ... ([5], p. 91).
>
> ...we must not hold that it is by means of this resemblance that the picture causes us to perceive the objects, as if there were yet other eyes in our brain with which we could apprehend it... ([5], p. 101).

This is René Descartes in his *Dioptrique* of 1637.

IV WHAT IS TACIT KNOWLEDGE ABOUT?

It is instructive to indicate some of the arguments with which Kosslyn has attempted to deflect the alternative tacit knowledge explanations of the much discussed experiments. Thus, for example, Kosslyn explains that 'virtually nobody outside the field knows about the McCollough effect' and therefore 'subjects could not have intentionally or unconsciously produced these results *without using imagery*' ([11], p. 81). There are two serious confusions in this response which illuminate the way in which the tacit knowledge argument has been misunderstood.

Firstly, the issue cannot be whether the subjects *used* imagery, since this is not in dispute. The question is only how best to explain the phenomena in which subjects *are, on all accounts,* using imagery. It is for *this* question that the tacit knowledge account is an alternative to the pictorial medium theory. That is, tacit knowledge purports to *explain* imagery and not to deny it. So it can be no argument against tacit knowledge to demonstrate how subjects *not* using imagery give different results from those who do. This has been a persistent mistake in the arguments offered by Kosslyn and provides the clearest indication of the manner in which imagery gets to be implicitly defined as pictorial by *fiat*.

A second serious error in the defence against tacit knowledge concerns Kosslyn's observation that subjects could not have known about such esoteric experimental results as the McCollough effect. Kosslyn suggests that certain properties of imagery 'are only manifest in highly novel laboratory settings, which precludes the subject from developing tacit knowledge about them from prior experience' ([12], p. 239). However, the relevant prior experience for acquiring tacit knowledge is not with the laboratory experiments, but with the normal use of the visual apparatus itself. Thus, the undoubted ignorance of subjects concerning such experimental results is entirely irrelevant to the possibility of explaining them through tacit knowledge. Pylyshyn's notion is paradigmatically exemplified in Chomsky's conception of our knowledge of grammatical rules. Accordingly, the appeal to tacit knowledge in the explanation of imagery is an appeal to rules and representations which are largely inaccessible to the subject and are used automatically in the relevant behaviour. Moreover, in the present context, the appeal to tacit knowledge is to invoke what the subject knows about the *phenomena* being imagined, namely the properties and regularities in the world—'a tacit physical theory' ([18], p. 201), and emphatically not about the arcane psychological experiments or their results as Kosslyn seems to think.

V CONFLICTING EVIDENCE

My own experiment [20] provides a direct test of the quasi-perceptual, pictorial medium account of imagery through testing the ability of subjects to rotate, scan and inspect the properties of their images and to make shape classifications etc, as has been widely claimed. I have specially designed the stimuli in order to test the idea that there can be a rigid angular transformation of a gross iconic image which is supposed to be inspected in order to discover new interpretations.

The stimulus figures are recognizable as a certain object in one orientation (1a), but are interpretable as an entirely different object when rotated 90 degrees (1b). It is important that the task of re-interpretation can be readily accomplished in a *perceptual* condition: When subjects are shown the figure in one orientation it is immediately recognized (as, say, the rabbit); upon rotating the

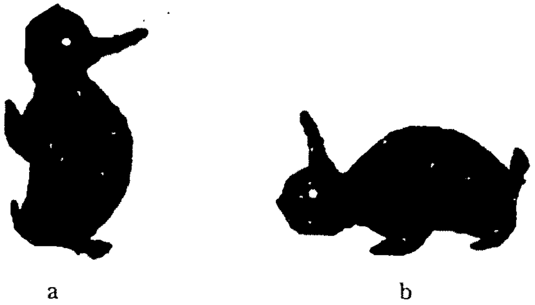


Figure 1.

figure by 90 degrees, subjects immediately discover (with frequent expressions of surprise and delight) the alternative interpretation (in this case the duck). Significantly, the figures are considerably simpler than the representations of blocks stacked in three dimensional shapes employed by Shepard and so it would be expected that they will pose no greater difficulty for mental rotation. Thus, subjects were shown the figures in only one orientation and then asked to *imagine* them rotated by 90 degrees. On the pictorial-medium, quasi-perceptual account, it would be expected that subjects could rotate their image and, by inspecting their rotated image, discover the new interpretation. However, contrary to what would be expected on the pictorial medium account, subjects were generally unable to discover any alternative interpretation in this way from their rotated image.

Thus, my preliminary investigations with figures of this kind have reinforced the negative findings of Chambers and Reisberg [2] and Rock et al. [19] using an experimental paradigm which has the important virtue of remaining very close to those of Shepard, Kosslyn and others. Moreover, it is essential to notice that the classical results of Shepard and Kosslyn were elegant linearities in reaction times, but such chronometric data are only very tenuously connected with any hypothesis concerning the nature of the underlying mechanisms or intrinsic properties of the substrate. In particular, and contrary to Kosslyn's consistent claims, these data are in principle incapable of supporting the pictorial medium account over Pylyshyn's tacit knowledge alternative since both accounts purport to explain the same reaction time data. The particular interest of the present experiment derives from its reliance upon a perceptual organization task which has an all-or-none character. Failure to discover an alternative interpretation of subjects' rotated images provides an unequivocal refutation of the expectations derived directly from the pictorial-medium theory: if images are raw, uninterpreted geometric shapes in a 'spatial' medium which can be inspected by the 'mind's eye', subjects ought to notice the novel alternative interpretations. On the other hand, if the images are highly interpreted conceptualizations as Pylyshyn and Clark have argued, then the rotated image preserves its original interpretation and does not permit re-parsing or the application of shape classification procedures in order to read off new properties. Under perceptual conditions, subjects readily notice the new construal of the figures such as the duck (or penguin); but when pressed under imagery conditions to re-interpret their rotated image, just as we would expect if the image is intrinsically bound to its interpretation, many subjects volunteer the response that it is a 'duckling on its back'!

Centre for Cognitive Science,
University of New South Wales,
Kensington, N.S.W. 2033 Australia

REFERENCES

[1] G. Berkeley, 'The Principles of Human Knowledge' in D. M. Armstrong, ed., *Berkeley's Philosophical Writings* (New York, Collier MacMillan, 1965).
[2] D. Chambers and D. Reisberg, 'Can Mental Images Be Ambiguous?', *Journal of Experimental Psychology: Human Perception and Performance*, 11 (1985) 317-28.
[3] A. Clark, 'Superman; The Image', *Analysis* 46 (1986) 222-4.
[4] A. Clark, 'Superman and the Duck/Rabbit: A Reply to Gordon and Bringsjord', *Analysis* 48 (1988) 54-57.
[5] R. Descartes, *Discourse on Method, Optics, Geometry and Meteorology*, translated by P. J. Olscamp (New York: Bobbs-Merrill, 1965).
[6] R. A. Finke, 'Levels of Equivalence in Imagery and Perception', *Psychological Review* 87 (1980) 113-32.
[7] R. A. Finke, *Principles of Mental Imagery* (Cambridge, Mass.: MIT Press, 1989).
[8] R. A. Finke, S. Pinker and M. J. Farah, 'Reinterpreting Visual Patterns in Mental Imagery', *Cognitive Science* 13 (1989) 51-78.
[9] R. Finke and K. Slayton, 'Explorations of creative visual synthesis in mental imagery', *Memory and Cognition* 16 (1988) 252-57.
[10] S. M. Kosslyn, *Image and Mind* (Cambridge, Mass.: Harvard University Press, 1980).
[11] S. M. Kosslyn, *Ghosts in the Mind's Machine* (New York: Norton, 1983).
[12] S. M. Kosslyn, 'The Medium and the Message in Mental Imagery: A Theory', *Psychological Review* 88 (1981) 46-66. Reprinted in N. Block ed., *Imagery* (Cambridge, Mass.: Bradford/MIT Press, 1981).
[13] S. M. Kosslyn, 'Seeing and Imagining in the Cerebral Hemispheres: A Computational Approach', *Psychological Review* 94 (1987) 148-75.
[14] S. M. Kosslyn, S. Pinker, G. Smith and S. P. Schwartz, 'On the demystification of mental imagery', *Behavioural and Brain Sciences* 2 (1979) 535-48.
[15] S. Pinker, 'Visual Cognition: An Introduction', in *Visual Cognition*, ed. S. Pinker (Cambridge, Mass.: MIT Press, 1985) 1-63.
[16] S. Pinker and R. Finke, 'Emergent two-dimensional patterns in images rotated in Depth', *Journal of Experimental Psychology: Human Perception and Performance* 6 (1980) 224-64.
[17] Z. Pylyshyn, 'Imagery and artificial intelligence', in *Minnesota Studies in the Philosophy of Science*, ed. C. Wade Savage, Vol. IX (University of Minnesota Press, 1978).
[18] Z. Pylyshyn, 'The Imagery Debate: Analog Media versus Tacit Knowledge', *Psychological Review*, 88 (1981) 16-45. Reprinted in N. Block ed., *Imagery* (Cambridge, Mass.: MIT Press, 1981).
[19] I. Rock, D. Wheeler and L. Tudor, 'Can we imagine how objects look from other viewpoints?, *Cognitive Psychology* 21 (1989) 185-210.
[20] P. Slezak, 'Rotating and Inspecting Images: A Crucial Experiment', in preparation.

EXTERNALISM AND EXPERIENCE

By GREGORY MCCULLOCH

EXTERNALISTS with respect to aspect of mind *A* say (roughly) that *A* ain't in the head; the corresponding internalists deny this. Colin McGinn has argued for what he calls 'weak externalism' concerning *experience*, illustrating the intended contrast thus: a state with content *That is F* may constitutively contain instances of *F*ness (strong externalism) – or it may merely contain the property *F*ness (weak externalism) – see [2] pp. 58–99; unattributed references are to this. But if externalism only holds of exerience in the weak form, then my (say) *having an experience as of sliminess* is no more (and no less) an external fact about me than my brain's *being* slimy. And in a perfectly ordinary sense, my brain's being slimy is an internal fact about me, made true by what is in my head. So McGinn's position on experience is internalistic in this same ordinary sense.[1] I have reservations about his distinction in so far as I deny that predicates stand for properties: but I waive that here. Instead, I argue first that his considerations apply to a much more limited area of experience than he makes clear. Second, I argue that he barely succeeds even here.

I

Imagine me enjoying a veridical visual experience of a square object. Now imagine my *doppelgänger* in a squareless world seeing a *round* object, but remaining atom-for-atom with me as I see the square object, because light reflected from the round object has been scrambled. Then, says McGinn, the *doppelgänger* has the same type of experience as me. If so, the content of such a state would not require the existence of square things, and at best weak externalism would hold here. He has two arguments. One is that if the content of experience were determined by the type of object confronted, introspective knowledge would be problematic. For example, if I were unknowingly switched to my *doppelgänger*'s world, then strong externalists would find it hard to avoid calling this a move from enjoying a square-thing-type experience to enjoying a round-thing-type experience (pp. 82–9; more on this later). But—since I could remain atom-for-atom as I was—nothing would show that I noticed: yet any account of such experiences should entail that such changes would be detectable by the subject. McGinn's other argument concerns the *doppelgänger's* behaviour. Since he is atom-for-atom with me he will move like me. And my movements will be appropriate to confronting a square: thus I might go squarewise around the object. But then so will my *doppelgänger* when skirting *his* object. This sits ill with the attribution to him of a round-thing-type experience: why does he make those right angle turns?

---

[1] McGinn seems to acknowledge the literal or ordinary sense in which weak externalism is internalistic. Having raised the question whether our thoughts about the mental states of others are externalistic, he claims that 'weak externalism is as true here as it is for other kinds of general concept...' (p. 52). He then goes on to object to strong externalism in this area, before concluding 'Given the essential role of one's own case for mental concepts, it looks as if internalism is going to find a secure niche in our account of them; weak externalism is the only kind of externalism we are going to get here' (p. 58). Elsewhere, however, he uses phrases like 'internal' and 'in the head' in a curious metaphorical sense: see p. 248 below.

McGinn concludes that 'strong externalism is not true of perceptual content' quite generally (p. 99). But consider this reply. If a content is perceptual if it can be ascribed intelligibly in the likes of 'X perceives that—', then there is little that is not perceptual. People perceive that daffodils, dogs, dynamos...are thus and so. But in many of these cases McGinn's two arguments do not apply. We are not introspectively infallible on whether we are experiencing dogs, and would not necessarily notice switching from dog-seeing to fake-dog-seeing. Moreover, if I see a dog and my *doppelgänger* sees a fake-dog, then although we move the same, this is no argument that we behave the same. There can be same movements, different behaviour (you sign a cheque, sipping H_2O; I give my autograph, sipping XYZ); and also different movements, same behaviour (we end the affair, you writing a letter, me telephoning). So my *doppelgänger* and I might move the same but behave differently, I stalking dogs, he stalking fake-dogs.

One may reply in turn that although I see a dog while my *doppelgänger* sees a fake-dog, still we both have experiences 'as of' a dog. McGinn himself uses the 'as-of' locution, and also claims to be concerned with what is '*given* in experience' (p. 59), with 'conscious seemings, states there is something it is like to have' (p. 63). And these matters are allegedly independent of questions of veridicality (pp. 58–9). Moreover, he distinguishes between 'observational' concepts (p. 59), that is 'concepts that can be applied...on the basis of experience without collateral information' (ibid.), and concepts whose application requires one 'to bring a bit of theory to bear' (p. 98).

But there is much unclarity here. Consider Hook, Line and Sinker. Hook sees that a dog is present in that her experience represents the scene as containing a dog. Line is similarly confronted by an identical dog, but innocent of most things zoological, has an experience which does not represent the scene thus. Finally Sinker, wise in the ways of dogs, sees a fake which she takes for real. There are various differences and similarities among these experiences. Perhaps they are 'qualitatively identical' (more of this below). Perhaps they only occur in situations containing dogs or good fakes. But here similarities run out. Given Line's innocence of dogs and suchlike, the beliefs and behaviour that result will be very different from the upshots in the cases of Hook and Sinker: and this grounds a sense in which Hook and Sinker *do* have experiences 'as of' a dog, whilst Line does *not*. For it is misleading to say that Sinker sees that a dog is present, since this carries at least some implication that there is a dog she sees: and that is false. So the term of art 'has an experience as of a dog' nicely captures the similarity between Hook and Sinker. Then although there is no parallel bar to saying that Line sees a dog, this is potentially misleading given her innocence: she does not see *that a dog is present*, since unlike Hook and Sinker, she does not see what she sees *as* a dog. As McGinn would say, her experience does not 'represent how things stand [with respect to dogs] in the external world' (p. 58). So if we want to mark the important difference between Hook and Line, and the correlative similarity between Hook and Sinker, it is perverse to describe Line as having an experience as of a dog.

McGinn remarks that 'One does not, strictly, have experiences as of water (say), but only as of colourless, tasteless liquid', and he would presumably say something similar about dogs. He also claims to be discussing 'subjective phenomenology' (p. 97). But one can resist the idea that Line enjoys the same phenomenology as Hook, whilst insisting that Sinker does, by claiming that the concepts which inform an experience actually affect its 'subjective phenomenology'. McGinn overlooks this, so that his use of this phrase begs an important question against strong externalism. For 'experiences as of' as just described are quite congenial to strong externalists, since the account makes having experiences as of dogs parasitic on having the concept of a dog, which is a concept of which strong externalism holds (if of anything). Nothing in McGinn's procedure rules out this kind of position, and so he has not really shown that strong externalism is false of 'as of' experience.

What of his mention of 'observational' concepts: could he not use this notion to develop his own sense of 'experience as of'? But the notion is elusive, as used by McGinn and otherwise. His phrase 'without bringing theory to bear' suggest theory-ladenness: seeing that a dog is present is only possible for those who know enough about dogs and related matters. So *dog* is not observational in McGinn's sense. But we learn in first-year philosophy of science that theory-laden terms appear in observation reports made by scientists. Thus 'there's a Great Dane in my study', 'there was copper in it', and so on. So these terms are 'theory laden' *and* observational: but not in McGinn's sense, presumably, since strong externalism holds of them (if of anything). Elsewhere he distinguishes between what is '*given* in experience' (p. 59, emphasis original), and what is inferred from this (along with 'collateral information'). However, this too is problematic, and moreover cuts across the point about theory-ladenness. Theory-ladenness can figure in experience: one can just see, without inference, that there is a Great Dane in the study. And if there are any non-theory-laden terms, then they can certainly be detected on the basis of an inference: supposing 'elliptical' to be such a term, Kepler still needed indirect methods to detect the shape of the Earth's orbit (cf. p. 98). It is evidently false that a concept is theory-laden as used by S if and only if S bases judgements containing that concept on an inference: indeed, both halves are false.

This is all tediously familiar, of course, but nevertheless it makes McGinn extremely difficult to interpret. The only other help he offers are references to Fodor and Peacocke (in [1] and [3] respectively; p. 59 n. 73). But Fodor is expressly discussing input systems, rather than perception and experience *per se* ([1] e.g. pp. 46, 73, 86–8). And when he ventures some 'highly speculative suggestions' ([1] p. 86; cf. n. 31) about what our visual input system delivers into consciousness, he says that

> in one useful sense of the observation/theory distinction, dogs but not protons count as observed...[and] phenomenologically accessible ([1] p. 97).

Peacocke says that a concept C is observational iff (1) there is some concept C' such that the capacity to have experiences as of something C', and to display a suitable sensitivity of one's judgements to the occurrence of such experiences, is needed for the possession of the concept C; (2) it would not be possible to have experiences as of something C' without having the concept C; and (3) C does not have the same pattern of epistemic possibilities as some concept which *fails* to meet (1) and (2) (where concepts C^* and $C\#$ have the same pattern of epistemic possibilities iff for any Σ, it is epistemically possible that $\Sigma(C^*)$ iff it is epistemically possible that $\Sigma(C\#)$) ([3] p. 94).

Applying this to get actual cases is a bit tricky (he suggests that the concept of a square meets it, whereas the concept of a tomato and the concept of a cause fail). But what he seems to be after is a 'primitive level of representational content' that our experience incorporates ([3] pp. 101–2). And I suggest that it might be helpful here to gloss this notion, which seems to be driving McGinn too, in the following way. Even if it is infelicitous to say that Hook, Line and Sinker all have experiences as of a dog, they have 'qualitatively' identical experiences. That is, quite independently of the above point that phenomenology may be shaped by informing concepts, we can say that experience has a qualitative or phenomenal nature to convey which requires the use of a limited number of (as I shall say) 'structural' concepts: typically—where visual experience is concerned—colour and shape concepts. Thus we might say that Hook, Line and Sinker all enjoy (or could come to appreciate) the same type of *visual field* (a distribution of shapes and colours), quite regardless of whether there is a richer level of phenomenology at which informing concepts play a part. Line does not see her dog as having a different shape and colour from Hook's simply because she does not see it 'as a dog'.

This rough but intuitive idea need not involve the view that all knowledge is inferred from seemings so construed. One merely claims that there is a structural dimension of similarity among experiences which empiricists noted but then expected to carry more weight than it can. Nor need one suppose that this similarity is 'non-conceptual' (cf. p. 62), whatever that means. For we have said that colour and shape concepts are at least closely involved with the structural. I suggest, then, that McGinn be seen as arguing for weak externalism concerning *such* perceptual contents, even though he does not exactly make this clear. But even on this ground, I shall now argue, he establishes rather little.

II

Taken literally, the externalist's 'the mind ain't in the head' means that the mind is not in the skull: hand me a brain and you do not hand me a mind. But McGinn takes 'in the head' metaphorically, including everything about subjects which does not involve the rest of the world (pp. 7–9). This gives two sorts of internalism: old-fashioned type-type identity theorists are strong internalists, behaviourists are weak internalists (as long as their account of behaviour does not contain ineliminable environment-involving descriptions). It is not clear whether McGinn is a weak or a strong internalist. He is happy to leave it open that a 'brain in a vat...can have all the weakly external mental states we have' (p. 105) — and this is strongly internalistic — although he seems to qualify this with behaviouristic ruminations (pp. 69, 106), which themselves then seem to be qualified in favour of the vat-brain (p. 106 n. 121). Not only is McGinn imprecise over observationality, but he is evasive over the vat-brain and thus over the extent of his internalism. But however this may be, does he refute strong externalism about structural experiential contents?

We can ignore the mere 'intuition' that a vat-brain would embody experiences structurally identical to those of a subject with an atom-for-atom brain, since our concern is with McGinn's *doppelgänger* arguments designed to support it. But there is a distinction which I ignored earlier. A *doppelgänger* case may involve switching — imagining someone unknowingly moved to a similar world — or it may be 'full-blooded', and involve imagining *doppelgänger* who inhabit similar worlds. Now in a switching case one can claim that mental states may be 'carried over' from the original world. I do not stop liking my dog because confronted with his *doppelgänger*, and if I were switched to Twin Earth, where the stuff in rivers etc. is XYZ, still I could, initially anyway, be thinking about H_2O when using 'water', and not XYZ, *even though* I would have been thinking about XYZ had I been born there. Even internalists, as long as they accept that states of mind have semantic properties (albeit inessentially), would find it hard to deny the parallel point, since it is simply a point about reference. But in so far as we imagine a switching case, and in so far as the *doppelgänger* carries over states had in the original world, then clearly McGinn's appeals to introspection and behaviour will not work, since the alleged square-thing-type experience in the new world would be parasitic on the subject's former dealings with squares, and strong externalism would be correspondingly untouched.

Naturally one might consider a full-blooded case from the start. But now we need to know more about the world envisaged, and the allowed interactions with it. The seen objects had better be out of reach, for otherwise it would be harder to maintain that the behavioural reactions of the *doppelgänger* justified the internalist verdict (cf. p. 64). The subjects had better not be credited with the concept of a square either, on pain of begging a crucial question. But this makes the appeal to introspection problematic, because the appropriate verbal reports would be harder to interpret. Since we are not trading in intuitions, the claim that no difference would *register* becomes much harder to support. Still, I suppose we can imagine the seen objects to be in the distance, our subjects to be embedded in mountainsides, and, though such as never to have encountered a square before, still able to make the ascription of experiences of shapes uncontentious, by talking of corners, numbers of sides and so on. Then, I think we must accept, McGinn is right to claim that the *doppelgänger's* experience is square-thing-type (structurally speaking), and strong externalism with respect to this dimension of mind is thus refuted.

But very little to suggest that the structural dimension of experience is in the head, literally or metaphorically, follows from this. No conclusion follows about vat-brains since they are not in the argument: the trick is to be turned by keeping the original subject fixed and varying the *doppelgänger* environment. McGinn suggests that vat-brains can be brought in as follows. If the structural aspects of experience are at least metaphorically in the head—i.e. environment-independent—then either they are literally in the head or they are constituted by behavioural/functional dispositions. But if the latter, then such contents may still be attributed to appropriate vat-brains if 'we just make the dispositions conditional on having a body (of a certain sort)' (p. 106 n. 121). The first problem is that no argument is given for thus making the dispositions conditional: and we are not concerned with intuitions about vat-brains, but with arguments.

There is a second, more serious problem anyway, which drastically reduces the scope of McGinn's conclusion. Our subjects' capacities to have square-thing-type experiences are grounded in their capacities to experience, and think about, *shapes generally*. Otherwise it would not be given that we were starting out with an uncontentious example of experiencing a square and then performing the relevant *doppelgänger* subtractions and manipulations. For recall the dialectical situation. McGinn's aim is to start with an uncontentious example of someone enjoying a square-thing-type experience — that is, an example acceptable to strong externalists — and then non-questionbeggingly to imagine someone sufficiently and relevantly similar who is, nevertheless, isolated from squares. And intuitions about vat-brains aside, can we really imagine a *doppelgänger* having (say) experiences structurally just like mine are here even though he lives in an environment which is entirely *empty* — devoid even of force-fields and light sources of appropriate contour and shape? Similarly, although we can no doubt imagine an environment lacking black things containing an atom-for-atom *doppelgänger* of me as I visually confront a certain poodle in standard conditions, can we also imagine this *doppelgänger* to live in a completely colourless environment?

I doubt that such possibilities can be coherently envisaged, and if this is correct, then it is clear that McGinn at most forces a small modification on the strong externalist. Whilst it is strictly false to say that the (structural) content of a black-square-type experience is only available in environments containing squares and black things, it remains assertible that the capacity for having experience even with just the structural features now under discussion is dependent on being embedded in an environment of shaped and coloured objects. Thus not even the structural dimension of experience is shown by McGinn to be in the head — metaphorically or otherwise. And putting this together with what has gone before, it is clear that strong externalism about experience generally survives McGinn's argument in only a very slightly weakened form.[2]

University of Nottingham,
Nottingham NG7 2RD

REFERENCES

[1] Jerry A. Fodor, *The Modularity of Mind* (Cambridge, MA: MIT Press, 1983).
[2] Colin McGinn, *Mental Content* (Oxford: Blackwell, 1989).
[3] Christopher Peacocke, *Sense and Content* (Oxford: Oxford University Press, 1983).

---

[2] Thanks to Tim Crane and Bob Kirk for helpful comments.

HE WANTS TO TRY

By ALFRED R. MELE

IS it a conceptual requirement on wanting to *try* to do something, *A* — or on an agent's *acting* on such a want — that the agent want to *A*? The cases with which I am familiar allegedly supporting a negative answer are, for a reason that I shall sketch, unconvincing. However, a negative answer is defensible, and partly by appeal to cases.

Consider the following suggestions. (1) Little Johnny might try to throw a ball over his father's head simply as a means of getting the ball to his father, without wanting at all to throw the ball beyond his father. (2) Sally might try to shatter her prized shatterproof widget in order to prove to a friend that the widget is shatterproof, while not being at all desirous of shattering it. Both agents, it may be alleged, *want to try* to perform the actions in question, in which case they want to try to do something, *A*, and act on that want, even though it is false that they want to *A*.

Given the agents' respective purposes, however, one may justifiably be reluctant to grant that Johnny and Sally are trying to do the deeds mentioned. Rather, it seems, Johnny, in an attempt to get the ball to his father, is behaving as he would behave if he *were* trying to throw it over his father's head, and Sally is acting as she would (or might) act if she *were* trying to shatter the widget. Arguably, what they are trying to do is precisely what they want to do — to get the ball to father and to prove that the widget is shatterproof — and their efforts incorporate their acting *as though they were trying* to throw the ball over father's head and to shatter the item (cf. H. McCann, 'Rationality and the Range of Intention', *Midwest Studies in Philosophy* 10, 1986, n. 20). And, of course, if they are not trying to perform the latter deeds, there is no reason to suppose that they *want* to try to perform them.

Hugh McCann has committed himself to the view that '*S* intends to try to *A*' entails '*S* intends to *A*'. The commitment rests partly on the contention that, because trying is not itself a species of action, the intention to try to *A* is 'totally vacuous' unless it is 'supplemented by a more substantive intention' (op. cit. p. 203). 'Trying', for McCann (p. 201), 'is a word that is general in meaning: it signifies the business of going about the performance of the action, and it is used when we have occasion to distinguish this enterprise from that of actually carrying the action off'. One might take essentially the same line about wanting: Because trying itself is not a kind of action, one can want to try to *A* only if one has a want that can give substance or determinate content to the want to try; and, for the same reason, an agent's *acting upon* a want to try to *A* requires that he have a supplementary want.

Even if this line is correct, however, must the supplementing want be a want to *A*? More importantly, granting that trying is not a species of action, is the claim about substance or content true?

Suppose that Brett promises to pay Belton fifty dollars if Belton tries to solve a certain chess problem within five minutes. It is a mate-in-two problem, a kind of problem that Belton has found very difficult; and Belton thinks it very unlikely that he will solve it. Brett assures Belton that he need not actually solve the problem, but he insists that Belton must *try* to solve it (in five minutes time), and not just pretend to be trying. To forestall the objection that Belton will want to solve the problem, because only his discovering the solution can guarantee that Brett will deem him to have tried, let us suppose that Belton is convinced that Brett is psychic (or omniscient) and can tell whether he has tried, independently of Belton's succeeding. Now, if and when Belton does try to solve the problem, there is a sense in which his mental activity is directed toward discovering the solution, the way to achieve checkmate in two moves. Belton's consideration of various possible moves must be guided by the thought of bringing a piece into a mating position. His consideration of chess moves will be organized around mating the king. In some sense, then, mating the king – which *is* solving the puzzle (provided it is accomplished in two moves) – is a *goal* in Belton's thinking about chess moves when he is trying to solve the problem. Still, I see no need to suppose that this goal must be *desired* (or wanted: I make no distinction) by Belton. Perhaps he is wholly indifferent to his actually mating the king. Nevertheless, wanting the fifty dollars, he wants to *try* to solve the puzzle; and he may act on that want.

What does Belton's trying to solve the puzzle amount to? Following McCann, we can say that it is 'the business of going about' solving the puzzle, where that is distinguishable from succeeding in solving it. Typically, when we engage in such business we want actually to perform the action, and in the absence of such a want we would devote our energies to other enterprises. But Belton has a reason for trying that is not also a reason for succeeding. He is convinced that he will be rewarded for engaging in the business of going about solving the puzzle (as we might say) quite independently of whether that enterprise issues in a solution. In such cases, an agent's trying to *A* does not depend upon his wanting to *A*.

Now, a desire to try *simpliciter* would be vacuous. Try *what*? But Belton's desire is a desire to try to *solve the puzzle*. The content of his desire approximates that of a desire to solve the puzzle; but it is such that the desire (not necessarily Belton) would be satisfied by an attempt – successful or not – to solve the puzzle, whereas the satisfaction of a hypothetical desire to solve the puzzle would require the agent's actually solving it.

It might be objected that desires to try to *A* – genuinely to try and not just to pretend to try – are functionally indistinguishable from desires to *A* and, consequently, that there can be no mere desires to try to *A*. But, setting the inference aside, the *premiss* is disputable. Arguably, a functional feature of a range of desires is that, partly in virtue of a particular desire's presence, the agent is disposed to experience satisfaction when the desire itself is satisfied (i.e. when what is desired is achieved). Desires to try to *A* can fill this role in situations in which desires to *A* would not. The example of Belton is a case in point, if he fails to solve the problem.[1]

Davidson College,
Davidson, NC 28036, U.S.A.

---

[1] I am grateful to John Heil, Hugh McCann, and Peter Smith for their comments on an earlier version of this paper.

CONSEQUENTIALISM AND THE UNFORESEEABLE FUTURE

By ALASTAIR NORCROSS

IF consequentialism is understood as claiming, at least, that the moral character of an action depends only on the consequences of the action, it might be thought that the difficulty of knowing what all the consequences of any action will be poses a problem for consequentialism. J. J. C. Smart writes that in most cases

> We do not . . . need to consider very remote consequences, as these in the end approximate rapidly to zero like the furthermost ripples on a pond after a stone has been dropped into it. . . . The necessity for the 'ripples in the pond' postulate comes from the fact that usually we do not know whether remote consequences will be good or bad. Therefore we cannot know what to do unless we can assume that remote consequences can be left out of account . . . If [such a postulate] is not accepted, not only utilitarianism, but also deontological systems like that of Sir David Ross, who at least admits beneficence as one *prima facie* duty among others, will be fatally affected. ([1], pp. 33–34)

Smart's claim seems to be that we needn't worry too much about the unforeseeable future, because most of our choices make very little difference to the state of the universe beyond what we can reasonably foresee. I do not see any good reason to believe Smart's 'ripples in the pond' postulate, but I also do not think that consequentialism needs such a postulate.

Smart's worry seems to be that our ignorance of remote consequences will strip us of the ability to make rational choices on consequentialist grounds. After all, if my choice now will go on making a difference in the world far beyond what I can foresee, how can I be justified in my belief that I am doing the right thing? I think that the consequentialist can admit that we are often faced with choices concerning which we have very little reason to believe that one choice will have better total consequences than the other. I also think that it is possible to make rational decisions in such cases on consequentialist grounds alone. Consider the following illustration: you are visiting a totalitarian dictatorship on a humanitarian mission, when the dictator himself, whom you have reason to trust, informs you that the fates of two political prisoners of conscience, Smith and Jones, are in your hands. You can specify whether Smith lives or dies. If you do not specify either fate for Smith, you will be taken to have specified death. In addition, you are told that the fate of Jones is tied to your decision about Smith. At this very moment, in a different part of the presidential palace, a coin is being tossed to link your decision about Smith to a decision about Jones. If the coin lands heads, life for Smith will also mean life for Jones, and death for Smith will mean death for Jones. If the coin lands tails, life for Smith will mean death for Jones, and death for Smith will mean life for Jones. You will not learn of the result of the coin toss, however. If your only concern is with the fates of Smith and Jones, it is clear what your choice must be. If you specify life for Smith, it will be certain that he will live, and it will be as likely as not that Jones will live too. If you specify death for Smith (even by a refusal to take part in the dictator's scheme), it will be certain that he will die, and it will be as likely as not that Jones will die too. You can, of course, employ rational choice theory, discounting the various possible outcomes by their probabilities. On this basis, a choice of life for Smith will have an expected outcome of one and a half lives saved; whereas a choice of death for Smith will have an expected outcome of half a life saved. If your only concern is with the fates of Smith and Jones, you must specify life for Smith.

What happens if we add a third prisoner, Wilson, whose fate is also tied to your decision about Smith by another coin toss? If you specify life for Smith, it will be certain that Smith will live, as likely as not that Jones will live, and as likely as not that Wilson will live (not as likely as not that both will live). Rational choice theory gives the result that a choice of life for Smith has an expected outcome of two lives saved, whereas a choice of death for Smith has an expected outcome of one life saved. Again, it is clear that you must choose life for Smith. Notice, however, that the three-person case has an important feature which the two-person case lacks. In the two-person case, a choice of life for Smith could not have a worse outcome than a choice of death for Smith. In the three-person case, it is possible, though unlikely, that a choice of life for Smith will result in fewer lives saved than a choice of death for Smith. This should not affect the rationality of your choice. If you have no more reason to think that a choice of death for Smith will benefit Jones and Wilson than a choice of life for Smith, and you know which will benefit Smith, you choose life for Smith. In this case, the unforeseeable consequences of your choice are greater than the foreseeable consequences, but the basis of your choice is still clear.

I do not think that your choice is any less clear when there are seven million other prisoners whose lives depend on your decision for Smith. Why would anyone think otherwise? There may be a temptation to regard one life as trivial when compared with seven million. What difference will a choice of life or death for Smith make when compared with the millions who will surely die whatever you choose? Or perhaps we could say that it is not so much that one more life is trivial compared with several million, but rather that morality should not have anything to say about such a difference. Bernard Williams could be taken to be describing such a view when he talks of a moral agent for whom 'there are certain situations so monstrous that the idea that the processes of moral rationality could yield an answer in them is insane: they are situations which so transcend in enormity the human business of moral deliberation that from a moral point of view it cannot matter any more what happens' ([2], p. 92). Williams contrasts such a view with consequentialism, which 'will have something to say even on the difference between massacring seven million, and massacring seven million and one' ([2], p. 93). One can certainly sympathize with the agent who is so horrified at the scale of a massacre that she finds it difficult to deliberate rationally in the circumstances. This does not, however, support the view that from a moral point of view it cannot matter anymore what happens. If there really is no moral difference between massacring seven million and massacring seven million and one, the allied soldier arriving at Auschwitz can have no moral reason for preventing the murder of one last Jew before the Nazi surrender. The Nazi himself can have no moral reason for refraining from one last murder. While Williams's moral agent is berating the universe for transcending the bounds of rationality, the consequentialist is saving a life. It is not hard to guess which of these agents I would rather have on my side.

The example of the difference between massacring seven million and massacring seven million and one is importantly different from my example of a choice of life or death for Smith tied to the fates of seven million other prisoners. Williams's example does not pose a problem of application for the consequentialist. It is clear that seven million and one deaths are worse than seven million. Rather, Williams's example is supposed to illustrate a feature of consequentialist reasoning which Williams finds disturbing. My example might be thought to be problematic for the consequentialist, because the agent has very little reason to think that a choice of life for Smith will have overall better consequences than a choice of death for Smith. I think, however, that there is an important similarity between the two cases. In each case there may be a temptation to abandon consequentialist methods of reasoning on the grounds that what one is able to bring about, or able to be certain to bring about, is somehow dwarfed by the scale of what else is beyond one's control, or beyond one's epistemic reach. In each case, the cure is to focus on what is within one's control. Whether one person lives or dies is always morally significant, indeed it is this very significance which is at the root of our horror at the sort of massacres described by Williams. If one can decide whether Smith lives or dies, and there is no more reason to believe that life for Smith will have consequences for others which are worse than the consequences of death for Smith, the choice is just as clear in the seven million and one person case as in the two person case.

My example, of course, does not exactly model our epistemic situation with regard to the future. We do not operate only with subjective probabilities of one and of one half. A more complicated example could be devised as a variation on the situation I describe. My point would remain. It is rational for a consequentialist to accept that a choice with foreseeable good consequences may, for all she knows, have unforeseeable bad consequences which massively outweigh the good consequences, and yet to make the choice purely on the basis of foreseeable consequences.[1]

Syracuse University,
Syracuse, NY 13244-1170, U.S.A.

REFERENCES

[1] J. J. C. Smart, 'An Outline of a System of Utilitarian Ethics', in *Utilitariansim For and Against* by J. J. C. Smart and Bernard Williams (Cambridge: Cambridge University Press, 1973).

[2] Bernard Williams, 'A Critique of Utilitarianism', in *Utilitarianism For and Against.*

[1] This paper has benefitted from discussions with Jonathan Bennett, Joel Kidder, Steven Lee and Rick Wiley.

UTILITARIANISM AND THE 'PUNISHMENT' OF THE INNOCENT: THE GENERAL PROBLEM

By SAUL SMILANSKY

OPPONENTS of utilitarianism have traditionally claimed that utilitarians might recommend the 'punishment' of innocent people (people who have not done anything wrong). And utilitarians have worked hard to deny this claim, both sides rightly perceiving the importance of this issue. For if utilitarianism does imply that innocent people ought sometimes to be 'punished', this is *prima facie* a most damaging feature of the utilitarian position.[1] The debate has been however conducted in a somewhat unreal atmosphere: anti-utilitarians have by and large come up with extreme and peripheral examples, which are often unlikely or even bizarre. And utilitarians have naturally seized on this, claiming that the problem of the 'punishment' of the innocent is not a *real* problem for utilitarians (the classical exchange is [4], [5], [6], vs. [9]; see also [10], p. 13f.). Both sides have seemed to share the opinion that in daily life (at least in the West) utilitarianism will not entail much (if any) serious transgression of the intuitively very strong moral ban on the 'punishment' of the innocent. I will argue that this basic notion is mistaken: in the creation and daily application of the criminal law we are constantly facing a general situation in which utilitarians would be obliged to promote the 'punishment' of the innocent.

The discussion between utilitarians and their opponents has typically proceeded as follows. The utilitarian is called to consider a case such as one where if a Sheriff were to frame an innocent person this would be beneficial, preventing a riot by a lynching mob. The utilitarian argues that (say) the Utilitarian Sheriff would not have the necessary knowledge about the probable course of events that would allow him to take such a decision, or that the act of injustice would become known and harm public trust, and the like. The anti-utilitarian then modifies his example to overcome the specific point made by the utilitarian, laying himself even more open to the claim that the problem is only manufactured and unreal. In day to day life, the utilitarian will tend to say, we want there to be well brought up Sheriffs (or judges or doctors etc., as the case may be) who are intuitively disinclined to be unjust. If we think critically as utilitarians, on the other hand, then perhaps in some fanciful cases acts of injustice would be right—but that after all says nothing about real life (see [1], p. 132f.).

The claim that the anti-utilitarians only appear to have a case if fanciful examples are being employed is common in utilitarian writing. Hare for example speaks about the 'vulgar objections...based on bizarre examples', with which 'A utilitarian theory of punishment can deal quite easily' ([2], p. 214). His typical claim against the anti-utilitarian objecting to utilitarianism because of the 'punishment' of the innocent issue is that 'The objection is usually based on highly artificial examples, for the good reason that no real ones are forthcoming which support the objector's case' ([3], p. 248). Similarly, J. J. C. Smart claims that '...a utilitarian might argue that it is empirically unlikely that some such situation as McCloskey envisages will ever occur...' ([8], p. 70).

---

[1] For the purpose of this paper the exact definition of utilitarianism and distinctions between forms of utilitarianism are not important. My arguments apply just as much to 'rule utilitarianism' as to 'act utilitarianism'. I discuss utilitarianism and not consequentialism because utilitarianism has been the main target of the 'punishment' of the innocent argument, and in order to focus the issue. The problem considered here would present a difficulty for most interesting forms of consequentialism, although there could of course be a consequentialist position in which the injustice involved would be considered bad as such, irrespective of other consequences.

Some anti-utilitarians have responded by saying that even if the utilitarian were right about 'real life', the fact that in principle utilitarianism might lead to abhorrent conclusions should in itself lead us to abandon it (e.g. [4], [5] pp. 254–55, [10] p. 18f.). And such a move is also naturally interpreted as reinforcing the utilitarian claim that in practice we have nothing to fear from the spread of utilitarian views.

I think that both types of anti-utilitarian moves — the presentation of detailed examples and the claim that the 'principle of the matter' is in itself important — are partially successful. It is complacent to think that utilitarians (critically deliberating as utilitarians) would not be lead in real situations to the conclusion that they ought to act in a way that would commonly be seen as gravely unjust. There is also a deep problem in any position which claims that there is no difference between (say) those 'blamed' or 'punished' properly (in common terms) and the framed innocent (or the 'punished' insane) — except that in the first case the general practical effects of 'blame' or 'punishment' are considered to be probably, on the whole, preferable. Some of the more realistic particular examples of the anti-utilitarians do create serious problems for the utilitarians, and even if we were to be tempted to agree that in practical terms the utilitarian outlook would not have harmful effects, this would still not make utilitarianism an acceptable theoretical position. However, the issue that I want to consider creates a difficulty for utilitarianism which transcends that made by the two sorts of issues just mentioned. The anti-utilitarian has a much stronger *general* case, which is very real and of central importance.

Take the current Western procedures for the apprehension of suspects, their prosecution and trial. And consider the suggestion that, perhaps even *primarily in order to save the innocent lives of victims of crime,* we should simply reduce the safeguards for conviction under normal trial procedures (so that fewer guilty criminals will get away). Procedures would remain much as they are today, no one ought to be framed, and a person's option of clearing his name by means of the common excuses (e.g. non-intentionality, ignorance or compulsion) would still exist. It would only be that matters such as the currently strong requirements as to the type of evidence needed for conviction, the degree of certainty which juries are instructed to take into account in their deliberations, the constraints on the types of evidence admissible in court, and the like, would be somewhat and gradually relaxed. Such a change is likely to have some effect on the number of cases that the prosecution will present to the courts and perhaps on policing as well. Why is such a suggestion a problem for utilitarianism? The reason is that with such a relaxation of procedures the following situation is likely to ensue:

1. Significantly fewer criminals would escape conviction, and (assuming the utility of punishment) this would benefit the vast majority of people: directly and indirectly it would reduce crime, the fear of crime, insurance premiums, policing and security costs, etc.
2. There would not be a significant increase in the arbitrary power of people in positions of authority.
3. General respect for the law and the judicial system need not diminish (even if the details of the procedures would be realized): as the judicial system will be significantly more effective, possibly just the reverse.
4. Most innocent people would still not be at any sort of probable risk of being 'punished'.
5. Most people's chances of being wrongly 'punished' might be negligible, compared to their chance of being severely hurt by a criminal who got away because of the previous stringent procedures.
6. More innocent people, however, are very likely to be 'punished'.

In other words, changing in the above ways the conditions under which it is decided whether people are to be punished would be positive in terms of a utilitarian 'cost analysis'. Only in terms of justice, where there is something inherently wrong in the 'punishment' of any innocent person, is such a transformation a bad thing. And so only a non-utilitarian (or not-only-utilitarian) position could guard us from such injustice. Such a position would insist that we are not simply to count numbers of lives likely to be saved or the balance of serious harm, under the various possible social conditions and procedures (even adding the particular suffering of those being innocently 'punished' and other such relevant utilitarian considerations). It would say that, as 'punishing' the innocent is in itself wrong, doing a great deal to avoid it is a matter of great importance — which must not be overridden almost irrespective of consequences.

The utilitarian is likely to have a number of replies. Firstly, she might doubt whether we can have such probable knowledge of the actual effects of the 'relaxation of standards'. This is a dangerous argument for utilitarians however, for utilitarianism in general depends on the strong possibility of having such knowledge of probable utilities, and excluding this in the case at hand would seem to be *ad hoc*. And it does seem reasonable to think that we can have the general sense of assurance about likely consequences which my case assumes. Perhaps it would be going too far to say that we can formulate a statistical law of 'increasing and diminishing utility', whereby relaxing standards at the price of risking more and more 'punishment' of innocent people would contribute a greater beneficial return (reducing crime), until we would reach a stage where the standards for conviction would be so lax as to cause 'too many' innocents to be 'punished' (i.e. an unacceptable number in utilitarian terms—a number that would create fear, lack of confidence in the judicial system, and the like). But even if we cannot formulate such 'laws', surely we have some working notion of the risks involved in various alternatives—after all, our current standards are widely perceived as practically protecting innocent people from being 'punished'.

Similarly, the utilitarian cannot get much mileage from the fact that the innocent people—those who are likely to be 'punished' if standards of criminal conviction are relaxed—are unknown, unlike in the case of framing an identifiable person. If the 'punishment' of innocent people is morally abhorrent than promoting it is so as well, and the anonymity involved seems not to matter.

A third possible utilitarian argument can be that in effect we are already involved in utilitarian calculations today, and are presumably willing to risk the 'punishment' of innocent people (or we would not have a punishing system at all). If this argument is simply a claim that even today we consider the likely outcome of various alternatives when determining policy, or even that there is a strong utilitarian rationale behind the institution of punishment in itself, then this point is harmless. However, it would be unconvincing to say that we currently have no strong non-utilitarian concern for preventing the 'punishment' of the innocent. Just consider the strength of the case that the prosecution has to make, and the strong provisions built into current trial (and appeal) procedures in order to help the accused respond to the prosecution's case (provisions made use of by many of the guilty in order to escape conviction). There is perhaps always a risk of injustice, but doing as much as we do to prevent it would seem to protect us from any description of quasi-utilitarianism.

A fourth and still different type of utilitarian response might be to embrace my argument and say that such balancing of risks is just what we should do: if (as I have claimed) it would be beneficial in utilitarian terms to relax current procedures in the above ways, then this is what ought to be done. Such a utilitarian reply would however be conceding my argument. For I am not directly arguing here for continuing to 'lose out' in terms of utility for the sake of justice: my claim is simply that the utilitarian claim, that in real life we have no *need to choose* between justice and utility, is unconvincing.

To conclude. The issue of the 'punishment' of the innocent is a very serious problem for utilitarianism. Both utilitarians and their opponents have perhaps not gone about considering this issue in the best way. Many anti-utilitarians have focussed on specific and often artificial examples. And utilitarians have tended not to tackle the crucial difficulty, seizing on the tactical mistakes of their opponents, instead of attempting to make a case for the consideration of utility when it contradicts justice (as it often does). For if we consider not specific detailed examples (or the radical transformation of social institutions)[2] but the utilitarian attractions of a limited relaxation of the rigorous critieria for the conviction of people on trial, within the current judicial framework — the conflict between considerations of utility and of justice is clear. There are strong utilitarian reasons for gradual and limited modifications such as would be very likely to result in grave injustice.[3]

Magdalen College,
Oxford OX1 4AU

REFERENCES

[1] R. M. Hare, *Moral Thinking* (Oxford: Oxford University Press, 1981).
[2] R. M. Hare, 'Punishment and Retributive Justice', in his *Essays on Political Morality* (Oxford: Oxford University Press, 1989).
[3] R. M. Hare, 'Some Reasoning about Preferences', in his *Essays in Ethical Theory* (Oxford: Oxford University Press, 1989).
[4] H. J. McCloskey, 'A Note on Utilitarian Punishment', *Mind* 72 (1963) 599.
[5] H. J. McCloskey, 'A Non-Utilitarian Approach to Punishment', *Inquiry* 8 (1965) 249–63.
[6] H. J. McCloskey, 'Utilitarian and Retributive Punishment', *The Journal of Philosophy* 64 (1967) 91–110.
[7] John Rawls, 'Two Concepts of Rules', in *The Philosophy of Punishment*, edited by H. B. Acton (London: Macmillan, 1969).
[8] J. J. C. Smart, in Smart & Bernard Williams, *Utilitarianism For and Against* (Cambridge: Cambridge University Press, 1973).
[9] T. L. S. Sprigge, 'A Utilitarian Reply to Dr. McCloskey', *Inquiry* 8 (1965) 264–91.
[10] C. L. Ten, *Crime, Guilt, and Punishment* (Oxford: Oxford University Press, 1987).

---

[2] My way of formulating the problem also overcomes another sort of utilitarian argument. In a well known article Rawls dismisses the issue of 'punishing' the innocent because it would involve setting up an alternative social mechanism (or 'institution') for carrying out such policy, and this mechanism itself would lack utility overall ([7], p. 111f.). But my point depends only on our familiar judicial process.

[3] I am grateful to Galen Strawson, to Hagar Kahana-Smilansky and to the Editor for comments on drafts of this paper.

MURPHY AND MERCY

By JACOB ADLER

THE conventional wisdom about mercy holds that mercy and justice are distinct and autonomous virtues. Jeffrie Murphy has recently called this conventional wisdom into question, arguing that mercy is either a vice or a part of justice:[1] If mercy requires a departure from strict justice, it is calling for injustice, a vice; if mercy never requires a departure from strict justice, it is then merely a part of justice (p. 169). The purpose of this paper is to acquit mercy of this charge.

The paper pursues the following plan. In Section I, I consider the nature of mercy and justice, and argue that mercy must be classified as a character-state or virtue. I next argue (Section II) that mercy is an autonomous virtue but (Section III) that it is nonetheless compatible with justice. Section III, in fact, describes two modes of the compatible interaction of mercy and justice. In Section IV – which is the crux of the paper – I show that mercy is distinguished from justice not only as a character-state, but also in respect of the results that these two virtues produce, and that mercy is needed if the right result is to be obtained: in other words, mercy is an autonomous judicial virtue.

Throughout this paper I consider mercy primarily in the judicial context, although, as Murphy notes (pp. 174–77), mercy has other uses.

I THE NATURE OF MERCY AND JUSTICE

To discuss our question sensibly we must first know a little about what mercy and justice are.

The word 'mercy' can be used to denote either a putative *virtue* (or character-state) or a *set of principles* prescribing certain behaviour. Correspondingly we may speak of a merciful *person* (one who has the character-state) or a merciful *action* (one that conforms to the principles). As a putative virtue, mercy has to do with the merciful person's character and motives, as Murphy describes quite aptly:

> As a moral virtue, mercy derives its value at least in part from the fact that it flows from a certain state of character – a character disposed to perform merciful acts from love or compassion while not losing sight of the importance of justice. (p. 166)

A person possessing the virtue of mercy must thus be able to feel compassion, to recognize these feelings, and to act on them; and she must be disposed to act on them. On the other hand, if mercy is a set of principles, all that one needs is a good will and an understanding of what behaviour the principles call for.

---

[1] Jeffrie G. Murphy, 'Mercy and Legal Justice', in Jeffrie G. Murphy and Jean Hampton, *Forgiveness and Mercy* (Cambridge: Cambridge University Press, 1988). Subsequent references thereto are indicated parenthetically by page number only.

Though Murphy is aware of this distinction,[2] he tends to speak of mercy only as a set of principles. For example, in one place he discusses mercy solely in terms of its results—namely, the choice of punishments—without reference to the character-state that produces those results (p. 171). Indeed, if mercy is just a set of principles, Murphy's conclusions are correct. Since the principles of mercy (so conceived) are just a particular subset of the principles governing punishment, it would be irrational to set them aside as an autonomous category distinct from justice. One might as well propose distinct virtues corresponding to the principles of excuses or of justifications.

If we do think of mercy as a character-state, we can delineate two ideal types: the just and merciful judge, and the just but non-merciful judge. For convenience I shall christen the former Judge Kirk, the latter Judge Spock. Kirk possesses the (putative) virtues of justice and mercy in an ideal degree. Spock is ideally just, but lacks mercy, and does not even think of mercy as a virtue. Nor does she think there are any distinct *principles* of mercy as contrasted with principles of justice. It is not to be thought that Spock is cruel or hard-hearted (as the word 'merciless' might suggest); she is simply dispassionate.

It should be noted that the compassion of the merciful judge must be disinterested. She feels compassion not because of some special relationship with the defendant, but merely because the defendant is a certain sort of person in a certain sort of difficult situation.

So much for the concept of mercy; next justice. Aristotle defines justice in one place as 'virtue entire...with regard to our neighbour' (*NE* 1129b27). This definition would make Murphy's thesis trivially true—presumably not what he intends. (Unfortunately, at least one of Murphy's arguments tacitly invokes this broad definition, as will be seen near the end of Section III.) Murphy must be using some narrower sense of the word. Although he gives no definition, Murphy characterizes justice as dealing with what each person is entitled to insist upon as a matter of right (pp. 166, 171–72), and I shall concur with this characterization.

II MERCY AS AN AUTONOMOUS VIRTUE

This characterization enables us easily to show that mercy is an autonomous virtue. For a merciful judge by definition feels compassion and acts on these feelings. But what each person is entitled to by right is in no way dependent on how other people feel about it. On the contrary, an entitlement is something that others must respect whether then feel inclined to or not. Hence is is a principle of justice that a judge has a prima facie duty to be dispassionate, to avoid being influenced by feelings she may have towards the litigants before her. A judge motivated by justice alone will therefore tend to set aside feelings of compassion. Mercy and justice, considered as states of character, are thus quite distinct.

---

[2] He states this most clearly in an earlier version of the paper. See Jeffrie G. Murphy, 'Mercy and Legal Justice', *Social Philosophy and Policy* 4.1 (1986), p. 11, n. 22.

The merciful judge, then, uses a different method of adjudication than her non-merciful colleague. This does not of course, prove that they will reach different results — that will be shown in Section IV. In the meantime we must consider whether mercy and justice are compatible.

III MERCY COMPATIBLE WITH JUSTICE

I shall assume for the moment that mercy will at least sometimes lead the merciful Judge Kirk to reach a different verdict than Judge Spock. The question then arises: Do not Kirk's merciful acts make her count as unjust, at least to a certain extent? The answer, in short, is No. In this section I present two models of the way in which these two virtues can compatibly interact: in the first model they are *cooperative*, in the second, *competitive*. These two models are not mutually exclusive; I consider it likely that any acceptable theory of mercy and justice would combine both.

First the *cooperative* model. We start with a suggestion of Murphy's. He argues that mercy — when not a vice — is merely a department of justice that calls for individuation:[3] the merciful judge individuates between those who deserve leniency and those who do not (pp. 170–72). Mercy is thus just a special case of the unquestioned principle of justice,

(1) Treat relevantly different cases differently. (p. 171)

To get mercy, we must, however, add another principle,

(2) The different degrees of disinterested compassion that a judge may feel towards a litigant are relevant differences.

A judge who accepts these principles will *want* to act mercifully; but it is only the virtue of mercy that enables her to perceive the relevant differences and to act accordingly. Judge Spock may well accept these two principles, but she will be unable to act on them.

---

[3] I follow Murphy here in using the verb 'to individuate' and its derivatives in their ordinary-language sense of 'to distinguish' or 'to individualize'. The rather different meaning of this term that is familiar from philosophical logic is, of course, not at issue here.

She would be like a colour-blind person trying to sort paint chips according to colour.

Thus, on the cooperative model justice tells us what we must do, while mercy enables us to do it.

Of course, we have also taken it to be a principle of justice that a judge's feelings towards litigants are irrelevant. Principle (2) must therefore either override the principle of judicial dispassion, or else constitute an exception. But certainly any reasonable theory of justice would allow a judge to pay attention to feelings of disinterested compassion if it could be shown that she would be more likely to reach a just verdict.

Now, I have just admitted that mercy is required by justice. One might object that mercy is then only a part or aspect of justice, not an autonomous virtue. But it is not so. The exercise of justice requires mercy, courage, wisdom, and indeed all the virtues; but that does not reduce them all to aspects of justice, unless we adopt the broad definition — already rejected — of justice as virtue entire.

Note that on this cooperative model, mercy is subordinate to justice. By contrast, according to the *competitive* model, mercy and justice are of equal status, and considerations of mercy on some occasions simply outweigh considerations of justice.

Now, one might think that injustice would result from this process. But it is not necessarily so. For it is the person who has all the virtues, not just one or two, who sets the standard for right deliberation and action. If the former behaves differently from the latter, it is the latter who is deficient. In particular, if Judge Kirk imposes a lighter sentence than Judge Spock, it is Kirk who is right.

But one might object: Even if it is conceded that Kirk is right, he is tolerating *injustice* in the interest of mercy.

There are two lines of response. First, it is enough to concede that Kirk is acting rightly, even if he tolerates some injustice (in the narrow sense); for he is completely just in the broad sense. And, secondly, in some situations the other virtues do not direct us to tolerate some injustice, but show us that justice is inapplicable. Aristotle's paradigm of this is friendship (*NE* 1155a25–27). If a friend insists, in the name of justice, that the gifts she gives her friends must match to the penny the gifts she receives from them, it is a case of excessive and inappropriate justice: true friends do not make an exact accounting, nor do we consider them unjust if they accept a degree of inequality. It is thus at least conceivable that mercy may function similarly, and that Judge Kirk may sometimes let mercy supersede justice without thereby introducing injustice. It remains to be shown, in the next section, that such cases actually occur.

Hence we have two ways of conceiving the interaction of justice and mercy as distinct but compatible virtues.

IV THE NEED FOR MERCY

In the previous section I have been assuming that Judge Kirk will judge differently than Judge Spock. In the present section I argue that this is actually the case. I do so by showing why mercy is needed, and why justice alone cannot fill this need. The present section thus addresses the crux of the issue.

Murphy's arguments here are impressive. He points out that much talk of mercy is founded on an 'overly restricted and simplistic conception of justice' (p. 172, n. 7), perhaps a mechanical application of *lex talionis*. Such a conception would often lead a judge to sentence offenders too harshly. Perceiving this harshness, the supporter of this simplistic conception invokes the concept of mercy to avert this unfair result. But (says Murphy) we need have no recourse to the concept of mercy for this purpose. An adequate conception of justice will be more complex, and will in particular call for finer individuation. For example, it is a matter of justice, not mercy, that we should treat a young, inexperienced criminal more leniently than a hardened career criminal, since the younger person is less responsible and consequently less culpable (pp. 170–71). I will argue, however, that the point cannot be generalized so as to eliminate mercy as an autonomous virtue. I present two arguments, one corresponding to each of the models of justice and mercy. In regard to the cooperative model, I show that the just but non-merciful judge will be *unable* to make certain necessary distinctions. In regard to the competitive model, I will show that she will consider herself *forbidden* by justice to make certain morally necessary distinctions.

First, the *cooperative* model. Punishment usually involves the administration of pain or suffering. Hence, in order to assign a just punishment, a judge must be aware of how much suffering a given sentence will cause. She must be able to tell when a defendant has suffered enough, when a sentence would be excessively cruel, and so on. And, as Murphy himself concedes, justice requires that these judgements be individualized: the judge must know not only how much suffering, say, a year in prison causes on the average, but how much suffering a year in prison X will involve for defendant Y. Now, how could a judge make such judgements? A firm grasp of psychological theory is necessary, but apparently not sufficient. One must be able to apply the theory to individual cases, which are almost infinite in variety. It seems that only a sense of empathy could let a judge know what sort of suffering a particular penalty will cause to this person. But empathy for the suffering of another is compassion, and compassion, when it motivates its subject to leniency, is mercy. Thus only a merciful person could distinguish and act on this relevant factor of suffering.

By contrast, the non-merciful Judge Spock will have only a dim awareness of how much suffering she is causing or relieving. If, for example, she has to assess an intense suffering of a kind that she is not familiar with, she is likely to underestimate it, with injustice resulting. Consequently, with regard to the cooperative model, Judges Kirk and Spock will differ not only in method but in result; and mercy is needed to reach the right result.

Next, the *competitive* model. Here the inflexibility stems from the conception of justice that we have joined Murphy in using: justice deals with what each person is entitled to insist upon, to demand as a right. When we are concerned with legal justice, this requirement is realized in part by adherence to the rule of law: citizens are not to be governed by extemporaneous decrees, nor by bills of attainder, nor by overly vague legislation, but rather by clear general principles announced in advance. Without the predictability that such principles provide, governmental action might be so capricious that citizens could not be said to have any effective rights at all. This requirement of generality and predictability necessarily limits the degree to which the law can individuate cases. It forbids the state from making some distinctions which justice would require if we could judge on a case-by-case basis.[4] Laws concerning prescription drugs may serve as an example. The law of most countries forbids selling certain drugs without a physician's prescription. It would have been more just, in one respect, if the legislature had individuated more finely by enacting a statute that said, 'Don't sell drugs to anyone who doesn't know what she's getting herself into'. This, however, would not only be unworkable, but unjust. A pharmacist would have no way of knowing when she may be incurring criminal liability. There is thus an avoidable tension between two requirements of justice: the call for individuation and the call for generality and predictability. It is thus not a failure of justice, but an inherent limitation, that the law must fall short of perfect individuation. Usually we tolerate such results, knowing that they are the best we can do; but sometimes these general principles will yield a result so oppressive that it is impossible to tolerate. In such cases, mercy (and allied virtues) come into play, allowing us to make individuations that justice would forbid; and only mercy allows us to reach the right result.

Now, one might suppose that second-order principles of justice could prevent such undesirable outcomes without having recourse to mercy. This solution would work in some cases; but clearly, the same problem would simply reappear at a higher level.

This process can be illustrated by the United States Constitution's prohibition of cruel and unusual punishment. Can this prohibition be a matter of justice alone? To a large extent, it *is* a matter of justice: principles of proportionality and of the fair distribution of burdens embody much of its force. But these principles must be given legal effect in the form of general principles known in advance. And we know that these general principles will on occasion permit a punishment that would be forbidden by ideally individualized justice. How do we know when to make exceptions to our general principles? Higher-order principles give some help, but in the end there seems to be nothing but mercy that can serve as a guide. The judge consults with her own sense of compassion, and thereby comes to know that a proposed punishment is too cruel.

---

[4] This is a commonplace of equal protection law in the United States. See, e.g., Gerald Gunther, *Cases and Materials on Constitutional Law*, Tenth Edition (Mineola, N.Y.: The Foundation Press, 1980), Chapter 10, Section 1.

Thus, mercy is required if a judge is to avoid doing grave wrongs. One might even say: Only a merciful judge can achieve the truly just result. If we concentrate on the result, it will seem that justice is the sole virtue involved; this seems to be what Murphy has done. If, however, we attend to the method by which this result is achieved, a greater complexity is revealed.[5]

University of Arkansas, 127 Kimpel Hall,
Fayetteville, AR 72701, U.S.A.

---

[5] My thanks to Jeffrie Murphy, Lisa Bivens, Margaret Holmgren, William Hay, W. J. Monagle, and the Editor and referee of *Analysis* for helpful suggestions.

NONBRANCHING AND NONTRANSITIVITY

By DOUGLAS EHRING

THEORIES of physical object identity over time specify a relation or relations *R* such that physical object stages are stages of the same object (i.e. are gen-identical) just in case they stand in relation *R* to one another. An analysis of identity can be demonstrated to lead to violations of transitivity by reference to any case in which more than one object stage at the same time stands in relation *R* to an earlier stage. One strategy for avoiding these violations of transitivity is to append a "nonbranching" restriction which excludes the possibility of multiple simultaneous stages' being gen-identical to the same earlier stage. In this paper, I argue that although this way of understanding the nonbranching condition does in fact immunize a theory of identity against violations of transitivity in the standard duplication case, this condition does not guarantee against all such violations. I propose a way of formulating the nonbranching condition which avoids this difficulty. I focus on the Causal theory of identity.

The Causal theory of physical object identity states that object stages s_1 and s_2 are stages of the same object only if these stages are causally connected. This causal condition must be supplemented with further conditions in order to provide a sufficient condition. The result will be some relation *R*, which includes a causal component, such that two stages are gen-identical if and only if they stand in relation *R* to each other. The Causal theory is vulnerable to the duplication objection if an object stage s_1 at t_1 may stand in relation *R* to object stages s_2 and s_3, both of which occur at a later time, t_2. The Causal theory will then require us to say that these later stages are stages of the same object as the earlier stage, and, thus, that these later stages, s_2 and s_3, are themselves gen-identical, given the transitivity of identity. But this conclusion is highly counterintuitive since s_2 and s_3 occur at the same time but differ in various respects, including location.

The counterintuitive conclusion can be avoided either by denying the transitivity of identity or by modifying the analysis. The first option of rejecting transitivity is the least appealing. The second strategy consists in adding a "nonbranching" condition to a causal analysis (see, for example, Chris Swoyer, 'Causation and Identity', in *Midwest Studies in Philosophy*, vol. 9, 1984, p. 603). One way of reading this restriction is as follows: s_1 at t_1 is a stage of the same object as s_2 at t_2 only if there is no stage at t_2 other than s_2 which stands in relation *R* to s_1. In the case at hand, we will not be forced to the conclusion that s_2 and s_3 are stages of the same object since s_2 (and s_3) will not qualify as gen-identical to s_1 given this nonbranching clause.

In order to demonstrate that this way of formulating the nonbranching clause does not rule out all violations of transitivity, we must recall some of the motivation for adopting the Causal theory over the rival Spatiotemporal Continuity account. Causal theorists have argued that their theory is better equipped than the Spatiotemporal Continuity theory to deal with at least two kinds of cases: spatiotemporally continuous but nonidentical objects, and temporally "gappy" objects. Shoemaker presents an example of the first type (see Sidney Shoemaker, 'Identity, Properties, and Causality', in his *Identity, Cause, and Mind*, Cambridge: Cambridge University Press, 1984, p. 241). A destroyer machine is set to destroy a table and a creator machine is set, independently, to create, at the same time, a qualitatively identical table from nothing. As a result, there comes to exist a series of spatiotemporally continuous table stages which are *not* all stages of the same table, contrary to Continuity theory. Since the earlier and later table stages are *not* causally connected, the Causal theory analysis of this case concurs with our intuition of non-identity. The second kind of case consists of gappy objects, i.e. objects which temporally go out of existence along with all their parts. Since the stages of gappy objects do not display spatiotemporal continuity, the Spatiotemporal Continuity theorist will incorrectly judge that these are definitely not stages of the same object. But since these temporally discontinuous stages may be causally connected the Causal theory does not rule out their gen-identity if the other conditions of the theory are satisfied.

The following general case, which borrows elements from both of these cases, demonstrates that the nonbranching condition, formulated to rule out only overlapping stages, does not protect the Causal Theory from all transitivity violations. A physical object stage may be linked to diverging causal paths even if there are no stages in any of these paths which occur simultaneously, i.e. in which there are no overlapping stages. This will be the case, for example, if a stage at t_1 gives rise to two causal paths, one of which follows a temporally continuous path beginning with t_1 and ending at a later time, t_4, and the other of which follows a gappy path leading from t_1 to t_5 (which is later than t_4). The t_5 stage has no direct causal connection to the t_4 stage. Since these stages are not *R*-related, the t_4 stage will not be a stage of the same object as the t_5 stage. However, given the Causal Theory and transitivity, we are forced to assert just this identity claim. The important point is that the nonbranching condition as formulated will not help in cases of this type since there are no overlapping stages.

A specific example of this general type will help illustrate. Suppose that there exists a truck at t_1 which undergoes the following sequence. On one causal path, from t_1 to t_4, all but one of the truck's parts gradually disappear, each for a temporary period. At each successive moment, as each of the original parts disappears, it is replaced by a new, qualitatively indistinguishable part, created coincidentally by a machine the operation of which is not causally linked to the truck at any earlier stage. At t_5, the last of the original parts as well as all of the new parts go out of existence permanently. Along another causal path, the parts which disappeared from t_1 to t_4 reappear, at t_5, in their original relations to each other, along with a new part which replaces the last original part, which is permanently extinguished at t_5. As a result of these two causal processes the following are true:

(1) At t_4, there exists one intact truck with one of the original parts along with replacement parts.
(2) At t_5, this truck disappears and is replaced with a truck which contains all but one of the original parts of the t_1 truck.

Between the two moments from t_1 to t_4, intuition, as well as the Causal Theory, tells us that the truck stages will be stages of the same truck. Identity across time is not threatened by gradual parts' replacement alone. It will follow given transitivity that the t_1 stage is a stage of the same truck as the t_4 stage. The Causal Theory will also entail that the t_1 and t_5 stages are stages of the same truck. Although these stages are not temporally continuous, they share most of the same parts. These stages are directly causally connected even if they display temporal gappiness. But if the t_1 and t_4 stages are gen-identical, as are the t_1 and t_5 stages, then given transitivity, so must the t_4 and t_5 stages be gen-identical. But in fact the t_4 stage is definitely *not* a stage of the same object as the t_5 stage. Consider, first, that these stages share no parts in common. Secondly, the later stage at t_5 does not derive from the earlier stage at t_4 by way of gradual parts' replacement. Finally, and most importantly for the Causal theory, the t_4 stage is not causally connected in any way to the t_5 stage. This last fact alone guarantees that the Causal theory will judge that these stages are not stages of the same object. But, as we have seen, transitivity considerations require that these stages be gen-identical. The nonbranching condition as formulated above will not protect the Causal theory from *this* violation of transitivity since in this example not more than one stage *at the same time* stands in relation *R* to an earlier stage.

What this type of case shows is that an adequate nonbranching clause must be formulated to exclude more than the possibility of simultaneous gen-identical stages (which are non-*R*-related). This case shows that given the possibility of gaps, there can exist branching causal paths, even though there are no simultaneous competing candidates for gen-identity with an earlier stage. The nonbranching rule should read as follows: *x* and *z* are gen-identical only if there does not exist some stage *y* such that *x* is *R*-related to *y* but *z* is not *R*-related to *y*, or such that *z* is *R*-related to *y* but *x* is not *R*-related to *y*. This version of the nonbranching rule is broader in scope than the earlier version considered and is sufficient to handle the kind of case described here.[1]

Southern Methodist University,
Dallas, Texas 75275, U.S.A.

---

[1] I would like to thank the Editor for his helpful suggestions.

ON OMNISCIENCE AND A 'SET OF ALL TRUTHS': A REPLY TO BRINGSJORD

By PATRICK GRIM

IN a number of pieces I have tried to present what I see as quite basic Cantorian problems for any set of all truths and for omniscience (see [4], [5], [6], [7]). In 'Grim on Logic and Omniscience' [2] Selmer Bringsjord suggests two escape routes. I think it can be shown that neither of these succeeds.

I

Bringsjord takes the power set axiom to be at the core of the Cantorian difficulties at issue, and the first escape he offers is simply to do without it: to adopt an alternative set theory such as ZF-Power in which such an axiom doesn't appear. 'Such a move vitiates, in one blow, the cottage industry that Grim has helped develop out of the power set axiom.' (p. 187)

This is not the first time such an escape has been suggested: in [13] Christopher Menzel proposed ZF-Power in particular as a solution for Cantorian problems plaguing a set of all truths appropriate to any given possible world.

Not too surprisingly, sacrifice of the power set axiom results in quite major technical limitations. At a single blow, for example, we are exiled forever from Cantor's paradise. Bringsjord, though fully aware of the major technical limitations at issue, thinks they're something we simply have to live with: an unfortunate cost of the need to maintain a set of all truths.

The truth of the matter, however—as I tried to indicate in my reply to Menzel ([5])—is that even sacrifice of the power set axiom isn't *enough* to escape Cantorian difficulties regarding a set of all truths. We'd have to sacrifice significantly more.

Here the argument is as follows.

Let us suppose we *did* have a set T of all truths within ZF-Power. Then by the axiom schema of separation, carried over directly from ZF, we would have as a subset of T a set of all truths satisfying a particular condition $B(x)$. As long as our basic language is rich enough to express the notion that a truth t is about a topic c, it appears, one such condition will be 'x is about a set of truths.' By separation, then, we would have a set C of all truths about sets of truths.

More formally, using 'Axy' to indicate that x is about y, our condition 'x is about a set of truths' is:

$$\exists y \forall z (z \in y \supset z \in T . \& \ Axy).$$

The axiom schema of separation, taken directly from ZF, is

$$\forall z_1 \ldots \forall z_n \forall a \exists y \forall x (x \in y \equiv x \in a \ \& \ B(x)),$$

where $z_1, \ldots, z_n$ are the free variables of $B(x)$ other than x, and the only restriction on $B(x)$ is that it does not contain y as a free variable (see for example [11], p. 175). Using T for a in this schema and our condition above for $B(x)$ would give us a set C of all truths about sets of truths:

$$\forall x (x \in C \equiv x \in T \ \& \ \exists y \forall z (z \in y \supset z \in T . \& \ Axy)).$$

The existence of a set C, however, would give us a reductio. C would, in particular, be larger than T. For consider any one-to-one function f from T into C, mapping truths onto truths concerning sets of truths, and consider further a set c' of all truths which do not belong to the sets their assigned element is about. Here using '$a(f(x))$' to indicate the set (or union of sets) $f(x)$ is about,

$$x \in c' \equiv x \in T \;\&\; x \notin a(f(x)).$$

Clearly C will contain some truth about c' just as it contains some truth regarding any set of truths. But by familiar reasoning f can assign no element of T to any truth about c'. C is larger than T.

One branch of the reductio, then, is this: given a set T of all truths, it appears, there would be a subset C larger than the set T of which it is a subset.[1]

Another branch is this. Each element of C is a truth. There are thus more truths than elements of T, and thus T cannot, as assumed, be a set of *all* truths.

As an escape from the Cantorian difficulties of C, then, Bringsjord's appeal to ZF-Power proves insufficient. Something above and beyond the power set axiom would have to be sacrificed. Perhaps additional restrictions on the axiom schema of separation are called for. Perhaps the language of the system must exclude or restrict the notion of truths about sets of truths—though since there clearly *are* truths about any sets of truths this would seem far from satisfactory. Or perhaps we should simply conclude that there is no set of all truths.

II

The second escape Bringsjord offers, with an eye to saving omniscience in particular, is to challenge the following definition of omniscience:

Df. 1 $\forall x(x \text{ is omniscient} \supset \forall p(p \text{ is true} \equiv K_x p))$.

That such a definition is worthy of challenge I have no doubt; as indicated in [3] this definition clearly won't do as an adequate characterization of a traditional God's knowledge, and for that reason I have standardly avoided it. One problem is this: that a being might qualify as omniscient on such a definition and yet hold any number of *false* beliefs. What such a definition requires is that all truths be known by an omniscient being, and that all things known by such a being be true. Since falsehoods believed by such a being are neither truths nor things known, however, such a definition puts no effective restrictions on the false beliefs of an omniscient being.[2]

---

[1] Note that this is not a problem for sets in ZF-Power in general: the argument relies essentially on the supposition that T is a set of all truths.

[2] In order to qualify as omniscient and yet hold false beliefs, of course, a being B would have to hold contradictory beliefs, and would moreover have to know that his beliefs *were* contradictory. Were such a being so flawed as to shamelessly

This is not however the challenge that Bringsjord presses against Df. 1. His attempt is rather to defend omniscience by proposing that it be redefined as follows:

Df. 2 s is omniscient $=_{\mathrm{df}} \forall p(\Diamond K_s p \equiv K_s p)$,

claiming that

> it's not unreasonable to maintain that there are certain true propositions God can't *possibly* know, such as what it's like to be ignorant, finite, shortsighted, etc...I think it's safe to say that there well *may* be propositional knowledge lurking here...it may very well be, for example, that I know that being a finite creature is like [...], where [...] can be filled in in such a way that a genuine proposition is denoted. ([2], p. 188)

One peculiarity here is of course that Bringsjord's defence of omniscience is on the grounds that there are true propositions which God does not and cannot know. Surely that alone would seem enough to compromise a traditional notion of omniscience.

There are several further problems that are bound to plague such an approach as well, however.

Putting the details of Bringsjord's definition aside for a moment, consider simply the kind of being he is trying to invoke here: a divine being which knows, well, not *everything*, but everything metaphysically suitable to such a being; a being which knows everything except, say, those propositions knowable only from the shortcoming, mess-making perspective of finitude.

A major difficulty is that even such a radically re-conceived 'omniscient' being would fall victim to the kind of Cantorian argument at issue. Here for convenience I will again phrase the argument in terms of a power set, though the lesson of section I is that it might also be phrased without it.

Consider the set of all truths that a divine being such as Bringsjord intends *would* know — the set T of all p such that $K_s p$, where we are careful to exclude supposedly propositional self-knowledge of finitude, shortcoming, mess-making, and the like.

Consider further all subsets of T, elements of the power set $\mathscr{P}T$. For each such subset there will be a distinct set-theoretical truth — a truth to the effect that a chosen truth T_1 *is* a member of that set, for example, or a truth to the effect that it is not. By Cantor's theorem, then, there will be more set-theoretical truths regarding membership in subsets of T than there are truths in T: there will be set-theoretical truths that T leaves out.

But set-theoretical truths are certainly *not* the kinds of shortcoming, mess-making propositions that are knowable only from a

---

recognize that he holds contradictory beliefs, however, he could nonetheless qualify as omniscient in the sense defined; nothing in the definition requires that an omniscient being avoid even blatantly obvious contradiction. For further discussion, however, see [9], esp. pp. 35–7.

perspective of finitude; these are precisely the kinds of truths all of which Bringsjord's being *would* be expected to know. The lesson of the Cantorian argument, however, is that given any set of all truths any being knows, there will be truths it does not—truths knowledge of which would moreover be perfectly appropriate the sort of being Bringsjord intends. If Cantorian difficulties pose a problem for a traditional notion of omniscience, then, they pose just as much of a problem for Bringsjord's. The escape fails.

Here the argument has been constructed in terms of the reconception of an omniscient being Bringsjord has in mind, however, rather than the one he strictly defines. Df.2 itself faces further difficulties.

Within Bringsjord's definition, God is intended to qualify as omniscient, even if he doesn't know all true propositions, on the grounds that he knows all that is possible for a being of his sort to know. Here Bringsjord quite consciously models his account on definitions of omnipotence in which God is claimed to be omnipotent on the grounds not that he can do literally anything, but that he can do all that is possible for a being of his sort to do. It's not particularly surprising, then, that Bringsjord's definition will also face many of the same difficulties that plague such definitions of omnipotence.

Consider in this light a being Necessary McIg, essentially such that he knows only that he is conscious. McIg, on the definition Bringsjord offers, would qualify as omniscient.[3]

Worse still, consider any being which is essentially such that it is non-conscious—here a boulder qualifies, perhaps, or tomato juice or the Pacific Ocean.[4] For any such being Bringsjord's definition would hold, and we'd be committed to the omniscience of tomato juice.

III

There *are* areas relevant to the Cantorian argument against a set of all truths or against omniscience, I think, that are well worthy of further critical scrutiny. One of these Bringsjord mentions in a footnote: the issue of set-theoretical semantics for propositional quantification.

[3] Necessary McIg is here patterned on Necessary McEar, a refinement in Bruce Reichenbach's [15] of an example that appears in Plantinga's [14]. Classics regarding this type of problem for omnipotence clearly include [10] and [12], and it continues to play a major role in for example [17] and [8]. A more sophisticated recent treatment of 'maximal omnipotence' in terms of an ability to accomplish 'shareable' tasks appears in C. Anthony Anderson's [1].

[4] Here I am obliged to L. Theresa Watkins for discussion. A similar problem posed for certain definitions of omnipotence by stones and other inanimate objects is noted by Swinburne in [16].

What Bringsjord claims to offer, however, are two ways of escaping the basic Cantorian argument. Neither of these succeeds.

State University of New York at Stony Brook,
Stony Brook, New York 11794-3750, U.S.A.

REFERENCES

[1] C. Anthony Anderson, 'Divine Omnipotence and Impossible Tasks: An Intensional Analysis', *International Journal for Philosophy of Religion* 15 (1984) 109–24.
[2] Selmer Bringsjord, 'Grim on Logic and Omniscience', *Analysis* 49 (1989) 186–9.
[3] Patrick Grim, 'Some Neglected Problems of Omniscience', *American Philosophical Quarterly* 20 (1983) 265–76.
[4] Patrick Grim, 'There is No Set of all Truths', *Analysis* 44 (1984) 206–8.
[5] Patrick Grim, 'On Sets and Worlds: A Reply to Menzel', *Analysis* 46 (1986) 186–91.
[6] Patrick Grim, 'Logic and Limits of Knowledge and Truth', *Noûs* 22 (1988) 341–67.
[7] Patrick Grim and Gary Mar, 'On Situations and the World: A Problem for Barwise and Etchemendy', *Analysis* 49 (1989) 143–48.
[8] Joshua Hoffman, 'Commentary on Edward Wierenga's "Omnipotence Defined"', presented at the 78th Annual Meeting of the American Philosophical Association, Eastern Division, December 29, 1981.
[9] Jonathan Kvanvig, *The Possibility of an All-Knowing God* (New York: St. Martin's, 1986).
[10] Richard R. La Croix, 'The Impossibility of Defining "Omnipotence"', *Philosophical Studies* 32 (1977) 181–90.
[11] Azriel Levy, 'The Role of Classes in Set Theory', in *Sets and Classes*, edited by Gert H. Muller (Amsterdam: North-Holland, 1976) 173–215.
[12] George I. Mavrodes, 'Defining Omnipotence', *Philosophical Studies* 32 (1977) 191–202.
[13] Christopher Menzel, 'On Set Theoretic Possible Worlds', *Analysis* 46 (1986) 68–72.
[14] Alvin Plantinga, *God and Other Minds* (Ithaca, N.Y.: Cornell Univ. Press, 1967).
[15] Bruce R. Reichenbach, 'Mavrodes on Omnipotence', *Philosophical Studies* 37 (1980) 211–14.
[16] Richard Swinburne, 'Omnipotence', *American Philosophical Quarterly* 10 (1973) 231–37.
[17] Edward Wierenga, 'Omnipotence Defined', *Philosophy and Phenomenological Research* 43 (1983) 363–75.

© *The copyright of each paper published in this volume of* ANALYSIS *is retained by its author(s)*